Dependent mixture models: clustering and borrowing information

Dependent mixture models: clustering and borrowing information

A. Lijoi[1], B. Nipoti[2] and I. Prünster[3]

[1] University of Pavia & Collegio Carlo Alberto, Italy
E-mail: lijoi@unipv.it

[2] University of Torino & Collegio Carlo Alberto
E-mail: bermardo.nipoti@unito.it

[3] University of Torino & Collegio Carlo Alberto
E-mail: igor@econ.unito.it

Abstract

Most of the Bayesian nonparametric models for non–exchangeable data that are used in applications are based on some extension to the multivariate setting of the Dirichlet process, the best known being MacEachern's dependent Dirichlet process. A comparison of two recently introduced classes of vectors of dependent nonparametric priors, based on the Dirichlet and the normalized σ–stable processes respectively, is provided. These priors are used to define dependent hierarchical mixture models whose distributional properties are investigated. Furthermore, their inferential performance is examined through an extensive simulation study. The models exhibit different features, especially in terms of the clustering behavior and the borrowing of information across studies. Compared to popular Dirichlet process based models, mixtures of dependent normalized σ–stable processes turn out to be a valid choice being capable of more effectively detecting the clustering structure featured by the data.

Key words and phrases: Bayesian Nonparametrics; Dependent Process; Dirichlet process; Generalized Pólya urn scheme; Mixture models; Normalized σ–stable process; Partially exchangeable random partition.

1 Introduction

Bayesian inference, either in parametric or in nonparametric form, is commonly based on the assumption that the observations $X_1, \ldots, X_n$ are drawn from a an exchangeable sequence of random elements $(X_i)_{i\geq 1}$. This means that, for any n, the distribution of the vector $(X_1, \ldots, X_n)$ is invariant with respect to permutations of its components. Such an assumption reflects an idea of analogy or homogeneity of the data and forms the basis for predictive

inference. Furthermore, it is nicely translated into a property of conditional independence and identity in distribution by virtue of the de Finetti representation Theorem, namely

$$\begin{aligned} X_i \,|\, \tilde{p} &\overset{\text{iid}}{\sim} \tilde{p}, \qquad i = 1, \ldots, n, \\ \tilde{p} &\sim Q, \end{aligned} \tag{1}$$

where $\tilde{p}$ is some random probability measure whose distribution Q plays the role of a prior for Bayesian inference.

It is apparent that exchangeability of observations is a strong assumption that fails in many problems of practical interest. This is the case, for instance, when the data originate from different studies or refer to experiments performed under different conditions: in such a context it is reasonable to preserve the homogeneity condition within data that are generated from the same study or experimental condition, while, at the same time, dropping the conditional identity in distribution for data emerging from different studies/experiments. Recent literature in Bayesian nonparametric inference has addressed this issue by proposing models that can accommodate for more general forms of dependence than exchangeability. Most of the proposals rely on the notion of partial exchangeability, as set forth by de Finetti (1938), that formalizes the above idea: although not valid across the whole set of observations, exchangeability can hold true within k separate subgroups of observations. Here, for ease of exposition and with no loss of generality, we confine ourselves to considering the case where $k = 2$. More formally, let $\mathbb{X}$ be a complete and separable metric space whose Borel σ–algebra is henceforth denoted as $\mathscr{X}$ and let $P_{\mathbb{X}}$ denote the space of all probability measures on $(\mathbb{X}, \mathscr{X})$. Introduce two (ideally) infinite sequences $X^{(\infty)} = (X_n)_{n\geq 1}$ and $Y^{(\infty)} = (Y_n)_{n\geq 1}$ of $\mathbb{X}$–valued random elements defined on the probability space $(\Omega, \mathcal{F}, P)$. The sequence $(X, Y)^{(\infty)} = (X_1, X_2, \ldots, Y_1, Y_2, \ldots)$ is termed *partially exchangeable* if, for any $n_1, n_2 \geq 1$ and for all permutations λ_1 and λ_2 of $\{1, \ldots, n_1\}$ and $\{1, \ldots, n_2\}$, respectively, the distributions of $(X_1, \ldots, X_{n_1})$ and $(Y_1, \ldots, Y_{n_2})$ coincide, respectively, with the distributions of $(X_{\lambda_1(1)}, \ldots, X_{\lambda_1(n_1)})$ and $(Y_{\lambda_2(1)}, \ldots, Y_{\lambda_2(n_2)})$. This notion is equivalently formulated as

$$\begin{aligned} &\mathbb{P}[X^{(\infty)} \in A^{(n_1)}, Y^{(\infty)} \in B^{(n_2)}] \\ &\qquad = \int_{P_{\mathbb{X}} \times P_{\mathbb{X}}} \prod_{i=1}^{n_1} p_1(A_i) \prod_{j=1}^{n_2} p_2(B_j)\, Q(\mathrm{d}p_1, \mathrm{d}p_2), \end{aligned} \tag{2}$$

for any $n_1 \geq 1$ and $n_2 \geq 1$, where $A^{(n_1)} = A_1 \times \cdots \times A_{n_1} \times \mathbb{X}^\infty$, $B^{(n_2)} = B_1 \times \cdots \times B_{n_2} \times \mathbb{X}^\infty$ with A_i and B_j in $\mathscr{X}$ for all i and j. Furthermore, Q, the de Finetti measure of $(X, Y)^{(\infty)}$, is a distribution of some vector $(\tilde{p}_1, \tilde{p}_2)$ of random probability measures (RPMs) on $\mathbb{X}$. Like in the exchangeable case (1), from a Bayesian perspective Q represents a prior distribution . In this framework, proposing a model for partially exchangeable observations is equivalent to specifying a distribution Q. A convenient definition of such a distribution should display a large topological support in $P_{\mathbb{X}} \times P_{\mathbb{X}}$ and a suitable degree of flexibility in describing a whole variety of dependence structures that range from independence of $\tilde{p}_1$ and $\tilde{p}_2$ to their almost sure identity, the latter corresponding to a Q degenerate on $P_{\mathbb{X}}$.

The first proposal of Q in (2) dates back to 1978 and appears in Cifarelli and Regazzini (1978), where a nonparametric prior for partially exchangeable arrays, defined as mixture of Dirichlet processes (DP), is defined. More recently, MacEachern proposed a general class of dependent processes (MacEachern, 1999) and defined a related dependent Dirichlet process (DDP) (MacEachern, 2000), which represented the seminal contribution for a large and highly influential body of literature. Reviews and key references can be found in Hjort, Holmes, Müller and Walker (2010). The use of these new classes of models has been made accessible also to practitioners by virtue of the development of suitable MCMC sampling techniques that allow to draw approximate posterior inferences. Furthermore, it should be mentioned that an R package, named *DP–package*, allows straightforward applications to a variety of dependent models. See Jara et al. (2011) for details. The present paper inserts itself in this line of research and its focus will be on a particular class of dependent RPMs that arise as mixtures of independent RPMs, where one component is common to all mixtures. This structure of dependence first appeared in Müller, Quintana and Rosner (2004), where vectors of RPMs were defined as mixtures of two DPs, one idiosyncratic and the other in common. More recently and still in the Dirichlet setting, in Hatjispyros, Nicoleris and Walker (2011) a multivariate Dirichlet process with a similar dependence structure has been considered and applied to the estimation of vectors $(f_1, \ldots, f_m)$ of densities, by resorting to a slice sampler. In Lijoi, Nipoti and Prünster (2013) a similar approach has been followed in a general setup: dependent RPMs are defined as normalization of dependent completely random measures, obtained as mixtures of one common and one idiosyncratic component. This approach leads to the definition of a whole class of dependent RPMs that turn out to be analytically tractable and amenable of use in applications.

As a matter of fact, most of the dependent RPMs used in applications

can be thought of as extensions to the multivariate setting of the DP. This is a natural choice, the univariate DP being a widely studied object with well known properties. Nonetheless, as shown e.g. in Ishwaran and James (2001, 2003); Lijoi, Mena and Prünster (2005, 2007a) for the exchangeable case, other choices for the nonparametric component are indeed possible and allow to overcome some of the drawbacks of the DP such as, for instance, its sensitivity to the total mass parameter and its simplistic predictive structure. See Lijoi, Mena and Prünster (2007a,b) for a discussion. Carrying out a comparative analysis of structural features of such models also in the multivariate setting is an important task, which, to the best of our knowledge, has not yet been addressed. Such an analysis, in addition to its practical implications, allows also to gain a deeper understanding of the inferential implications of the various modeling choices. This paper aims at giving a contribution in this direction by comparing a bivariate DP with a bivariate normalized σ–stable process. The analysis that is going to be developed relies on the construction proposed in Lijoi, Nipoti and Prünster (2013). Moreover, dependent DPs and normalized σ–stable processes are the natural candidates to compare since many quantities of interest can be obtained in closed form. The nature of our comparison will therefore be two–fold: on the one hand, we will analyze different properties of the two vectors of RPMs by investigating their predictive distributions, while on the other hand we will resort to a simulation study in order to appreciate the difference between the two models when applied to problems of density estimation and clustering. Importantly, the results of the simulation study find intuitive explanations by means of the insights gained on the predictive structures of the models.

The outline of the paper is as follows. In Section 2 we concisely summarize the vector of bivariate RPMs introduced in Lijoi, Nipoti and Prünster (2013). A description of the dependent mixtures and a sketch of the MCMC algorithm that is implemented for drawing posterior inferences is provided in Section 3. In Section 4 we compare the properties of bivariate Dirichlet and normalized σ–stable processes by investigating the structure of their predictive distributions and the distribution of the total number of clusters that both models induce on two vectors of observations. Finally, Section 5 is devoted to an extensive simulation study. The inferential impact of the two models choices and of their characterizing parameters is analyzed by focusing on the estimation of the number of clusters of the two distributions and on the mechanism of borrowing strength between different studies. A concise description of the implementation of the Gibbs sampler is provided in the Appendix.

2 Dependent normalized completely random measures

Many popular nonparametric priors Q for exchangeable data, as in (1), arise as suitable transformations of completely random measures (CRMs). See Lijoi and Prünster (2010) for a survey of various classes of discrete nonparametric priors using CRMs as unifying concept. Here we also consider models with CRMs as basic building blocks and then rely on the idea of Lijoi, Nipoti and Prünster (2013) for defining a distribution Q of vectors $(\tilde{p}_1, \tilde{p}_2)$ of dependent random probability measures by normalizing vectors of dependent CRMs. For this reason we concisely recall the notion of CRM, which also allows us to introduce the main notation used throughout.

Suppose $M_{\mathbb{X}}$ is the space of boundedly finite measures on $\mathbb{X}$ whose Borel σ–algebra is denoted as $\mathscr{M}_{\mathbb{X}}$. For details see Daley and Vere–Jones (1988). A CRM μ is a random element defined on $(\Omega, \mathscr{F}, \mathbb{P})$ and taking values in $(M_{\mathbb{X}}, \mathscr{M}_{\mathbb{X}})$ such that for any $A_1, \ldots, A_n$ in $\mathscr{X}$, with $A_i \cap A_j = \varnothing$ when $i \neq j$, the random variables $\mu(A_1), \ldots, \mu(A_n)$ are mutually independent. Any realization of a CRM is a discrete measure with probability one and, if no fixed jump points are present, then

$$\mu = \sum_{i \geq 1} J_i \, \delta_{Z_i}, \tag{3}$$

for some sequences $(J_i)_{i \geq 1}$ and $(Z_i)_{i \geq 1}$ of random elements taking values in $\mathbb{R}^+$ and $\mathbb{X}$, respectively, and δ_x is the unit point mass at x. A CRM as in (3) is characterized by the so–called Lévy–Khintchine representation, which provides an expression for the Laplace functional transform of μ. Indeed, there exists a measure ν on $\mathbb{R}^+ \times \mathbb{X}$ such that $\int_{\mathbb{R}^+ \times \mathbb{X}} \min(s, 1) \nu(\mathrm{d}s, \mathrm{d}x) < \infty$, and

$$\mathbb{E}\left[\mathrm{e}^{-\int_{\mathbb{X}} f(x)\mu(dx)}\right] = \exp\left\{-\int_{\mathbb{R}^+ \times \mathbb{X}} \left[1 - \mathrm{e}^{-f(x)s}\right] \nu(\mathrm{d}s, \mathrm{d}x)\right\}, \tag{4}$$

for any measurable function $f : \mathbb{X} \to \mathbb{R}$ such that $\int |f| \, \mathrm{d}\mu < \infty$ almost surely. The measure ν takes on the name of *Lévy intensity* and, by means of (4), it uniquely identifies the CRM μ.

In order to define a vector of dependent CRMs $(\tilde{\mu}_1, \tilde{\mu}_2)$, we draw inspiration from an approach set forth in Griffiths and Milne (1978), where a class of bivariate vectors of dependent and identically distributed Poisson random measures is introduced. In a similar fashion, we shall consider identically distributed CRMs $\tilde{\mu}_1$ and $\tilde{\mu}_2$, with the same Lévy intensity ν, defined

as suitable mixtures of three independent CRMs μ_0, μ_1 and μ_2. These are characterized by their respective Lévy intensities ν_0, ν_1 and ν_2

$$\nu_0 = (1 - Z)\,\nu, \qquad\qquad \nu_1 = \nu_2 = Z\,\nu,$$

for some random variable Z taking values in $[0, 1]$ and independent of μ_i, for $i = 0, 1, 2$. More precisely, we set

$$\tilde{\mu}_1 = \mu_1 + \mu_0, \qquad\qquad \tilde{\mu}_2 = \mu_2 + \mu_0. \tag{5}$$

Hence, each $\tilde{\mu}_\ell$ is characterized by an independent CRM μ_ℓ and by a shared one μ_0, which induces dependence. Besides having an intuitive interpretation, the dependence introduced in (5) is appealing since it leads to a joint Laplace transform for $(\tilde{\mu}_1, \tilde{\mu}_2)$ with a simple structure. This property is inherited by the proposal of Griffiths and Milne (1978) and the availability of an explicit expression for the joint Laplace transform is pivotal in proving the results achieved in Lijoi, Nipoti and Prünster (2013). We therefore refer to (5) as GM–dependent CRM.

Now, by means of a suitable transformation of a GM–dependent CRM $(\tilde{\mu}_1, \tilde{\mu}_2)$, we are in a position to define the mixing measure Q in (2). To be more specific, if $\mathbb{P}[\tilde{\mu}_\ell(\mathbb{X}) \in (0, \infty)] = 1$, for $\ell = 1, 2$, we shall consider the vector

$$(\tilde{p}_1,\, \tilde{p}_2) \stackrel{d}{=} (\tilde{\mu}_1/\tilde{\mu}_1(\mathbb{X}),\, \tilde{\mu}_2/\tilde{\mu}_2(\mathbb{X})) \tag{6}$$

and the RPMs $\tilde{p}_1$ and $\tilde{p}_2$ are also termed *GM–dependent*. Each $\tilde{p}_\ell$ admits an interesting representation as a mixture of two independent RPMs, namely

$$\tilde{p}_\ell = w_\ell p_\ell + (1 - w_\ell)p_0, \qquad \ell = 1, 2, \tag{7}$$

where $w_\ell = \mu_\ell(\mathbb{X})/[\mu_\ell(\mathbb{X}) + \mu_0(\mathbb{X})]$ and $p_\ell = \mathbb{1}_{(0,1]}(z)\mu_\ell/\mu_\ell(\mathbb{X})$ for $\ell = 1, 2$, while $p_0 = \mathbb{1}_{[0,1)}(z)\mu_0/\mu_0(\mathbb{X})$. Note that $\mathbb{1}_A$ denotes the indicator function of set A. The weights w_1 and w_2 are dependent and, in general, not independent of p_0, p_1, p_2. The random variable Z plays a crucial role. If $Z = 0$ (almost surely), then $\tilde{p}_1$ and $\tilde{p}_2$ coincide (almost surely) and Q degenerates on $P_{\mathbb{X}}$: such a condition yields exchangeability of $(X, Y)^{(\infty)}$. In contrast, if $Z = 1$ (almost surely) then $w_1 = w_2 = 1$ in (7) and, therefore, $\tilde{p}_1$ and $\tilde{p}_2$ are independent. Hence, the magnitude of random variable Z provides an indication of the distance between the actual dependence in $(X, Y)^{(\infty)}$ and exchangeability. Note that in the exchangeable case one obtains the class of priors introduced in Regazzini, Lijoi and Prünster (2003), whose main inferential properties were derived in James, Lijoi and Prünster (2006, 2009). Before describing mixture models governed by GM–dependent

RPMs, we analyze the two specific cases that will be the object of our analysis.

Example 1. (*GM–dependent Dirichlet processes*). In order to obtain a dependent Dirichlet vector, one starts by considering gamma CRMs whose Lévy intensity is given by

$$\nu(\mathrm{d}s, \mathrm{d}x) = \frac{e^{-s}}{s}\, c\, P_0(\mathrm{d}x),$$

for some $c > 0$ and some probability measure P_0 on $\mathbb{X}$. Henceforth it will be assumed that P_0 is non–atomic. By normalizing $\tilde{\mu}_1$ and $\tilde{\mu}_2$, as defined in (5), one obtains a vector $(\tilde{p}_1, \tilde{p}_2)$ whose components are two GM–dependent DPs, identically distributed with total mass c and mean measure P_0. We denote such vector by GM-$\mathscr{D}(c, z, P_0)$ or, simply, GM–$\mathscr{D}$. Moreover, as for the mixture representation of $\tilde{p}_1$ and $\tilde{p}_2$ in (7), we observe that p_0, p_1 and p_2 are independent DPs with common mean measure P_0 and total mass respectively equal to $(1 - Z)c$, Zc and Zc.

Example 2. (*GM–dependent normalized σ–stable processes*). Consider a σ–stable CRM, whose Lévy intensity is given by

$$\nu(\mathrm{d}s, \mathrm{d}x) = \frac{\sigma s^{-1-\sigma}}{\Gamma(1-\sigma)}\, P_0(\mathrm{d}x),$$

with $\sigma \in (0, 1)$ and P_0 being probability measure on $\mathbb{X}$ that will be assumed non–atomic. The normalization of $\tilde{\mu}_1$ and $\tilde{\mu}_2$, as defined in (5), yields a vector $(\tilde{p}_1, \tilde{p}_2)$ whose components are GM–dependent normalized σ–stable processes, identically distributed with parameter σ and base measure P_0. We denote such vector by GM–st(σ, z, P_0), or more concisely, GM–st. Note that in this case we have set the total mass as $c = 1$. This does not cause any loss of generality since for normalized σ–stable CRMs, the parameter c is redundant. With reference to the mixture representation of $\tilde{p}_1$ and $\tilde{p}_2$ in (7), we observe that p_0, p_1 and p_2 are independent normalized σ–stable processes with common parameter σ and base measure P_0. However, in contrast to the Dirichlet case, the weights (w_1, w_2) and the mixed RPMs (p_0, p_1, p_2) are not independent.

Finally, note that both $\tilde{p}_1$ and $\tilde{p}_2$ select discrete probability distributions on $(\mathbb{X}, \mathscr{X})$, almost surely. This implies that $\mathbb{P}[X_i = X_j] > 0$ and $\mathbb{P}[Y_i = Y_j] > 0$ for any $i \neq j$ so that ties occur with positive probability within each group of observations. Moreover, it will be henceforth assumed that $\mathbb{P}[Z < 1] > 0$, which, in turn, entails $\mathbb{P}[X_i = Y_j] > 0$ for any i

and j: there is, then, a positive probability of detecting ties also between groups. Such properties naturally lead to consider the random partition induced by the data and, then, determine the probability of observing a sample $\{X_1, \ldots, X_{n_1}\} \cup \{Y_1, \ldots, Y_{n_2}\}$ having a specific configuration or partition structure. This extends the analysis of the random partition in the exchangeable case (1), for which the *exchangeable partition probability function* (EPPF) is a key tool. Remarkably, a closed form expression for the more general partially exchangeable case, named *partially exchangeable partition probability function* (pEPPF), is available for any vector of GM–dependent RPMs (Lijoi, Nipoti and Prünster, 2013, Proposition 2). A simple investigation of such a pEPPF shows that exchangeability holds within three separate groups of clusters that are governed by the three independents RPMs p_0, p_1 and p_2. This invariance property will be better described, in the next section, in the context of mixture models and is the key ingredient in devising an algorithm that can be thought of as an extension of the Blackwell–MacQueen Pólya urn scheme.

3 Dependent hierarchical mixtures

One of the most widely used models in Bayesian Nonparametrics is a hierarchical mixture, where a random probability measure $\tilde{p}$ is used as a mixing measure. Indeed, if Θ is some complete and separable metric space and $h : \mathbb{X} \times \Theta \to \mathbb{R}^+$ a transition kernel such that $x \mapsto h(x, \theta)$ is a density function on $\mathbb{X}$, for any $\theta \in \Theta$, then

$$\tilde{f}(x) = \int_{\Theta} h(x, \theta)\, \tilde{p}(\mathrm{d}\theta) \tag{8}$$

defines a random density function on $\mathbb{X}$, whose probability distribution is a prior on the space of density functions. When $\tilde{p}$ is a Dirichlet process, then $\tilde{f}$ is the Dirichlet process mixture introduced by Lo (1984) and popularized thanks to the MCMC sampler proposed in Escobar and West (1995) that has made its use straightforward in applied problems.

Here we consider an extension of this model that accommodates for experiments yielding two groups of observations $\{X_1, \ldots, X_{n_1}\}$ and $\{Y_1, \ldots, Y_{n_2}\}$ generated from random densities $\tilde{f}_1$ and $\tilde{f}_2$. These are defined by

$$\tilde{f}_\ell(x) = \int_{\Theta} h(x; \theta)\, \tilde{p}_\ell(\mathrm{d}\theta), \qquad \ell = 1, 2. \tag{9}$$

Moreover, for $\ell = 1, 2$, let $\boldsymbol{\theta}^{(\ell)} = (\theta_{1,\ell}, \ldots, \theta_{n_\ell,\ell})$ stand for vectors of latent variables corresponding to the two samples. Then the mixture model can

be represented in hierarchical form as

$$\begin{aligned}(X_i, Y_j) \,|\, (\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}) &\overset{\text{ind}}{\sim} h(\,\cdot\,;\theta_{i,1})\, h(\,\cdot\,;\theta_{j,2}),\\ \theta_{j,\ell} \,|\, (\tilde{p}_1, \tilde{p}_2) &\overset{\text{iid}}{\sim} \tilde{p}_\ell, \qquad\qquad j = 1, \ldots, n_\ell, \quad \ell = 1, 2,\\ (\tilde{p}_1, \tilde{p}_2) &\overset{d}{=} \text{ GM–dependent normalized CRM.}\end{aligned} \tag{10}$$

As remarked in the previous section, the combination of the almost sure discreteness of $\tilde{p}_\ell$ and of the dependence structure introduced in (7) implies that there can be ties within each vector $\boldsymbol{\theta}^{(\ell)}$ and among elements of the two vectors $\boldsymbol{\theta}^{(1)}$ and $\boldsymbol{\theta}^{(2)}$ with positive probability. Therefore, for $\ell = 1, 2$, there will be $k_\ell + k_0 \leq n_\ell$ distinct values in $\boldsymbol{\theta}^{(\ell)}$ that we denote as

$$\{\theta^*_{1,\ell}, \ldots, \theta^*_{k_\ell,\ell}, \theta^*_{1,0}, \ldots, \theta^*_{k_0,0}\}. \tag{11}$$

In (11) $\theta^*_{i,1}$, for $i = 1, \ldots, k_1$, does not match any value of the other vector $\boldsymbol{\theta}^{(2)}$. On the other hand $\theta^*_{i,0}$, for $i = 1, \ldots, k_0$, is shared by both vectors $\boldsymbol{\theta}^{(1)}$ and $\boldsymbol{\theta}^{(2)}$. The description of the partition structure of the sample is then completed by the corresponding frequencies. For each $\ell = 1, 2$, we denote them by

$$\{n^*_{1,\ell}, \ldots, n^*_{k_\ell,\ell}, q^*_{1,\ell}, \ldots, q^*_{k_0,\ell}\}.$$

Moreover, for $i = 1, \ldots, k_0$, $n^*_{i,0}$ indicates the sum $q_{i,1} + q_{i,2}$, that is the frequency of $\theta^*_{i,0}$ when both vectors are considered.

The simulation algorithm we are going to use relies on the decomposition displayed in (7) and on two collections of (non observable) auxiliary random variables $\boldsymbol{\zeta}^{(1)} = (\zeta_{i,1})_{i\geq 1}$ and $\boldsymbol{\zeta}^{(2)} = (\zeta_{j,2})_{j\geq 1}$ such that $\mathbb{P}[\zeta_{i,1} = 1] = 1 - \mathbb{P}[\zeta_{i,1} = 0] = w_1$ and $\mathbb{P}[\zeta_{j,2} = 2] = 1 - \mathbb{P}[\zeta_{j,2} = 0] = w_2$. We can then provide an alternative representation of the mixing measure in (10) in terms of these auxiliary variables as

$$\begin{aligned}\theta_{i,1} \,|\, \zeta_{i,1}, \mu_1, \mu_2, \mu_0 &\overset{\text{ind}}{\sim} p_{\zeta_{i,1}}, \qquad\qquad i = 1, \ldots, n_1,\\ \theta_{j,2} \,|\, \zeta_{j,2}, \mu_1, \mu_2, \mu_0 &\overset{\text{ind}}{\sim} p_{\zeta_{j,2}}, \qquad\qquad j = 1, \ldots, n_2,\\ (\zeta_{i,1}, \zeta_{j,2}) \,|\, \mu_1, \mu_2, \mu_0 &\overset{\text{iid}}{\sim} \text{bern}\,(w_1; \{0, 1\}) \times \text{bern}\,(w_2; \{0, 2\})\,.\end{aligned} \tag{12}$$

Now observe that

$$P[\theta_{i,\ell} = \theta_{j,\kappa} \,|\, \zeta_{i,\ell} \neq \zeta_{j,\kappa}] = 0,$$

for $\ell, \kappa \in \{1, 2\}$ and $i = 1, \ldots, n_\ell$, $j = 1, \ldots, n_\kappa$, which means that the latent variables can coincide only if their corresponding auxiliary variables coincide. If the latent variables are associated to different groups, that is

$\ell \neq \kappa$, they can match only if the corresponding auxiliary variables are both equal to 0. Thus, the auxiliary variables corresponding to the distinct values appearing in $\boldsymbol{\theta}^{(\ell)}$ in (11) can be written as

$$\{\zeta^*_{1,\ell}, \ldots, \zeta^*_{k_\ell,\ell}, \zeta^*_{1,0} = 0, \ldots, \zeta^*_{k_0,0} = 0\},$$

To sum up, $\boldsymbol{\theta}^{(1)}$ and $\boldsymbol{\theta}^{(2)}$ can be gathered into three separate groups, U_0, U_1 and U_2, according to the values taken by the corresponding auxiliary variables $\zeta_{i,1}$ and $\zeta_{j,2}$. For $\ell = 1, 2$, elements in $\boldsymbol{\theta}^{(\ell)}$ that are labeled with ℓ will end up in group U_ℓ, while variables of both groups with label 0 will form U_0. The latter includes all variables that display ties between vectors. Each group U_i has cardinality $\bar{n}_i$ and consists of $\bar{k}_i$ distinct values, where

$$\bar{n}_1 = \sum_{i=1}^{n_1} \zeta_{i,1}, \quad \bar{n}_2 = \sum_{i=1}^{n_2} \zeta_{i,2}/2, \quad \bar{n}_0 + \bar{n}_1 + \bar{n}_2 = n_1 + n_2.$$

Moreover, $\bar{k}_1 = \sum_{i=1}^{k_1} \zeta^*_{i,1}$, $\bar{k}_2 = \sum_{i=1}^{k_2} \zeta^*_{i,2}/2$ and $\bar{k}_0 + \bar{k}_1 + \bar{k}_2 = k_0 + k_1 + k_2$. The distinct values appearing in group U_i, if $i = 0, 1, 2$, will be denoted by $\{\tilde{\theta}^*_{1,i}, \ldots, \tilde{\theta}^*_{\bar{k}_i,i}\}$ and the corresponding frequencies will be $\{\tilde{n}_{1,i}, \ldots, \tilde{n}_{\bar{k}_i,i}\}$.

Another important feature of GM–dependent RPMs is that it is possible to find a closed expression for the joint distribution of the observations $(X_1, \ldots, X_{n_1}, Y_1, \ldots, Y_{n_2})$, the latent variables $(\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)})$, the auxiliary variables $(\boldsymbol{\zeta}^{(1)}, \boldsymbol{\zeta}^{(2)})$ and possible further parameters, after integrating out the CRMs μ_0, μ_1 and μ_2: this will, then, allow us to derive all the full conditionals that are needed in order to implement the MCMC algorithm we are going to describe.

Our goal is to apply this model to estimate the density functions of the X_i's and of the Y_j's and the number of clusters each sample displays. This is achieved by devising a Gibbs type algorithm that features three independent Pólya urns and that can be summarized in the following steps:

1. generate initial values for the latent variables, the auxiliary variables and the parameters;
2. update the auxiliary variables and the parameters using their full conditional distributions;
3. divide the latent variables in three groups according to the values of the auxiliary variables;
4. update the latent variables of each group via independent Pólya urn schemes;
5. go back to step 2.

4 Prediction and clustering with GM–dependent Dirichlet and normalized σ–stable CRMs

Hierarchical mixture models with (almost surely) discrete mixing measure offer a flexible and effective framework for model–based clustering. In fact, they naturally induce a distribution on the number of clusters the data can be grouped in, which can then be estimated through the posterior distribution. Looking at the clustering structure naturally leads to studying random partitions. Various proposals of priors for random partitions for partially exchangeable data, which belong to or are allied to the class of product partition models of Hartigan (1990), have appeared in the literature. In these cases the prior is covariate dependent and a borrowing strength phenomenon, analogous to the one we are studying here, takes place. See, e.g., Leon–Novelo et al. (2012) and Müller, Quintana and Rosner (2011). See also Petrone and Raftery (1997). However, a comparative analysis of the inferential implications of the choices of the nonparametric mixing measures has not yet been carried out in the partially exchangeable setting. Here we fill this gap and specifically focus on the clustering behavior associated to hierarchical models of the type described in (10), which allows us to replace the Dirichlet process with alternative discrete mixing measures in a quite straightforward way. In particular, we devote the present section to the analysis of the predictive and related clustering properties associated to GM–$\mathscr{D}$ and GM–st normalized CRMs $(\tilde{p}_1, \tilde{p}_2)$. Posterior inferences on the clustering of the data are then determined by the marginal partition probability function induced by each $\tilde{p}_\ell$ on the latent variables $\boldsymbol{\theta}^{(\ell)}$, as in the exchangeable case, and by the specific dependence structure between $\tilde{f}_1$ and $\tilde{f}_2$ governed by the mixing measures. This second aspect is new and peculiar to the partially exchangeable scheme in (10) that is examined. Our analysis proceeds in two directions. In this section we study the clustering and dependence structure by investigating the predictive distribution induced by the mixing measures and the prior distribution of the total number of clusters that they induce. In the next section, relying on such findings, we will carry out an extensive simulation study.

In order to further simplify notation, henceforth we shall use the symbol $\mathbb{P}_z$ to denote a probability distribution conditional on $Z = z$ and $\mathbb{E}_z$ as the corresponding expected value.

4.1 Predictive structures

With reference to the model (10), one can, in line of principle, determine the predictive distribution for $\theta_{n_\ell+1,\ell}$, for $\ell = 1, 2$, given two vectors of observations $\boldsymbol{\theta}^{(1)}$ and $\boldsymbol{\theta}^{(2)}$ governed by any GM–dependent normalized CRM via the evaluation of the corresponding pEPPF. One could, then, handle both GM–$\mathscr{D}(c, z, P_0)$ and GM–st(σ, z, P_0) mixtures in a similar fashion as in the exchangeable case. Unfortunately the resulting expressions, although available in closed form, are not of immediate use since they involve sums that are hard to compute even for small sample sizes n_1 and n_2. See Nipoti (2011); Lijoi, Nipoti and Prünster (2013). Nonetheless these analytical results display a key invariance property, recalled in the previous section, that leads to devise a simple MCMC simulation algorithm. In particular, since exchangeability holds true within three separate groups of observations identified by the realization of the auxiliary variables $\boldsymbol{\zeta}^{(1)}$, $\boldsymbol{\zeta}^{(2)}$, one can, then, determine the predictive distribution of $\theta_{n_\ell+1,\ell}$, given $\boldsymbol{\zeta}^{(1)}$, $\boldsymbol{\zeta}^{(2)}$ and $\zeta_{n_\ell+1,\ell}$. The invariance property and the corresponding representation in terms of the, although non–observable, auxiliary variables $\boldsymbol{\zeta}^{(1)}$ and $\boldsymbol{\zeta}^{(2)}$ yields very neat interpretations of the underlying predictive structure. Also, the examination of the two extreme cases of independence between $\tilde{p}_1$ and $\tilde{p}_2$ (that corresponds to Z degenerate at $z = 1$) and almost sure identity (that is, Z degenerate at $z = 0$) provides further insight on the structural properties of the model. According to the framework described in the previous section

$$\mathbb{P}[\theta_{n_\ell+1,\ell} \in U_i \,|\, \zeta_{n_\ell+1,\ell} = j] = \begin{cases} 1 & \text{if} \quad i = j \\ 0 & \text{otherwise,} \end{cases}$$

where U_0, U_1 and U_2 denote, as before, three groups into which $\boldsymbol{\theta}^{(1)}$ and $\boldsymbol{\theta}^{(2)}$ are gathered together. Therefore, the conditional predictive distribution of $\theta_{n_\ell+1,\ell}$ for GM–$\mathscr{D}$ processes coincides with

$$\mathbb{P}_z[\theta_{n_\ell+1,\ell} \in \cdot \,|\, \boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \boldsymbol{\zeta}^{(1)}, \boldsymbol{\zeta}^{(2)}, \zeta_{n_\ell+1,\ell} = i] = \frac{\bar{z}c}{\bar{z}c + \bar{n}_i} P_0(\cdot) + \sum_{j=1}^{\bar{k}_i} \frac{\tilde{n}_{j,i}}{\bar{z}c + \bar{n}_i} \delta_{\tilde{\theta}^*_{j,i}}(\cdot), \quad (13)$$

where $i \in \{0, \ell\}$, $\bar{z} = (1-z)\mathbb{1}_{\{0\}}(i) + z\mathbb{1}_{\{\ell\}}(i)$. Recall that $\tilde{n}_{j,i}$ and $\bar{n}_i$ identify the number of components in $\boldsymbol{\theta}^{(1)}$ and in $\boldsymbol{\theta}^{(2)}$ equal to $\theta^*_{j,i}$ and belonging to group U_i, respectively. Similarly, for GM–st processes, we have

$$\mathbb{P}_z[\theta_{n_\ell+1,\ell} \in \cdot \,|\, \boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \boldsymbol{\zeta}^{(1)}, \boldsymbol{\zeta}^{(2)}, \zeta_{n_\ell+1,\ell} = i] = \frac{\sigma \bar{k}_i}{\bar{n}_i} P_0(\cdot) + \sum_{j=1}^{\bar{k}_i} \frac{\tilde{n}_{j,i} - \sigma}{\bar{n}_i} \delta_{\tilde{\theta}^*_{j,i}}(\cdot). \quad (14)$$

Interestingly the expressions in (13) and (14) correspond to the well–known predictive distributions of the DP and of the normalized σ–stable processes, respectively. This is due to the fact that, conditionally on the latent variables ζ's, the analysis of the predictive distributions induced by a pair of GM–dependent processes boils down to the study of the three (conditionally) independent processes, whose behavior is known. The mechanism for allocating the mass underlying the distributions in (13) and in (14) is best illustrated as the result of a two step procedure. The first step corresponds to the generation of either a new value $\tilde{\theta}^*_{\bar{k}_i+1,i}$ or of one of the already observed values $\{\tilde{\theta}^*_{1,i}, \ldots, \tilde{\theta}^*_{\bar{k}_i,i}\}$. In the GM–$\mathscr{D}$ case, the corresponding probability coincides with

$$\mathbb{P}_z \left[\theta_{n_\ell+1,\ell} \notin \{\tilde{\theta}^*_{1,i}, \ldots, \tilde{\theta}^*_{\bar{k}_i,i}\} \,|\, \boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \boldsymbol{\zeta}^{(1)}, \boldsymbol{\zeta}^{(2)}, \zeta_{n_\ell+1,\ell} = i\right] = \frac{\bar{z}c}{\bar{z}c + \bar{n}_i},$$

which depends solely on the number of observed values $\bar{n}_i$ and the total mass $\bar{z}c$. In contrast, for the GM–st case one has

$$\mathbb{P}_z \left[\theta_{n_\ell+1,\ell} \notin \{\tilde{\theta}^*_{1,i}, \ldots, \tilde{\theta}^*_{\bar{k}_i,i}\} \,|\, \boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \boldsymbol{\zeta}^{(1)}, \boldsymbol{\zeta}^{(2)}, \zeta_{n_\ell+1,\ell} = i\right] = \frac{\sigma \bar{k}_i}{\bar{n}_i}, \quad (15)$$

which depends explicitly on the number of observed clusters $\bar{k}_i$, in addition to $\bar{n}_i$ and the model parameter σ. Therefore the latter predictive structure is richer in that it makes explicit use of a larger portion of the sample information for dictating the allocation between new and already observed values. As for the second step in prediction, one has that, if $\theta_{n_\ell+1,\ell}$ is a new value, then it is sampled from P_0; instead, if $\theta_{n_\ell+1,\ell}$ is not new, one deduces the probability of coincidence with any $\tilde{\theta}^*_{j,i}$ from (13) and (14). Such coincidence probabilities are proportional to the cluster size $\tilde{n}_{j,i}$ in the GM–$\mathscr{D}$ case, while they are not proportional to the cluster size for the GM–st process, since they depend on σ as well.

These different predictive features are very influential when developing an analysis of the clustering of the data as we shall detail in the next subsection.

4.2 Prediction and clustering

Although in both considered models the predictive distribution is a convex linear combination of the base probability measure $P_0 = \mathbb{E}[\tilde{p}_\ell]$ and a possibly weighted empirical distribution, the resulting mass allocation among "new" and "old" distinct values in $\boldsymbol{\theta}^{(\ell)}$, for $\ell = 1, 2$, is significantly different therefore affecting posterior inferences on clustering.

First, in both cases the probability of observing a new value is an increasing function of the parameter of the process, that is c or σ: it is concave for the DP and linear for normalized σ–stable processes. More interestingly, in the GM–st setting, such a probability also depends on $\bar{k}_i$: this is an appealing feature since it formalizes the idea that, for a fixed sample size $\bar{n}_i$, the larger the number of already observed distinct values, the higher the probability of observing a new one. Hence, the larger $\bar{k}_i$, the larger the estimated number of clusters tends to be. As a further remark, specific to the dependent case, note that z appears, through $\bar{z}$, only in (13). This might lead, in the GM–$\mathscr{D}$ model, to a significant discrepancy between the probabilities of observing a new value in U_0 and U_ℓ. Indeed, if we suppose that $\bar{n}_0 = \bar{n}_\ell$, we have that

$$\frac{\mathbb{P}_z[\theta_{n_\ell+1,\ell} = \text{"}new\text{"} \mid \zeta_{n_\ell+1,\ell} = 0, \ldots]}{\mathbb{P}_z[\theta_{n_\ell+1,\ell} = \text{"}new\text{"} \mid \zeta_{n_\ell+1,\ell} = \ell, \ldots]} = \frac{1-z}{z}. \tag{16}$$

Hence, if $z < 0.5$, which means we are closer to a situation of total exchangeability where $\tilde{p}_1 = \tilde{p}_2$, then the ratio in (16) is greater than 1, while if $z > 0.5$, which corresponds to being closer to independence between $\tilde{p}_1$ and $\tilde{p}_2$, the same quantity is smaller than 1. For a fair analysis of this last feature, one has to take into account that this tendency is balanced by the fact that the parameter z plays also a role in determining the cardinalities $\bar{n}_0$ and $\bar{n}_\ell$ since, for both GM–$\mathscr{D}$ and GM–st models, $\mathbb{E}_z[w_1] = \mathbb{E}_z[w_2] = z$. In other words $\mathbb{P}_z[\zeta_{i,\ell} = \ell] = 1 - \mathbb{P}_z[\zeta_{i,\ell} = 0] = z$ for any i and ℓ.

The description of the predictive structure in GM–dependent models is then completed by describing the mass allocation to already observed values $\{\tilde{\theta}^*_{1,i}, \ldots, \tilde{\theta}^*_{\bar{k}_i,i}\}$. By looking at the ratio of the probabilities assigned to any pair of observed values $(\tilde{\theta}^*_{j,i}, \tilde{\theta}^*_{l,i})$, for a GM–$\mathscr{D}$ model this is equal to the ratio of their cardinalities $\tilde{n}_{j,i}/\tilde{n}_{l,i}$ and, therefore, each cluster is assigned mass proportional to its cardinality. Things are significantly different in the case of GM–st models, in which the parameter σ plays a key role. In fact, in terms of the probability of generating a new value displayed in (15), it is apparent that the larger σ, the higher is such a probability. Now, once a new value has been generated, it will enter the predictive distribution of

the next step: since it will clearly have frequency 1, from (14) one sees that its mass will be proportional to $(1-\sigma)$, instead of 1 as in the GM–$\mathscr{D}$ case, and, correspondingly, a mass proportional to σ is added to the probability of generating a new value. Therefore, new values are assigned a mass which is less than proportional to their cluster size (that is 1) and the remaining mass is added to the probability of generating a new value. On the other hand, if a value is re–observed, the associated mass is increased by a quantity which is now proportional to 1, and not less than proportional. This implies that the ratio of the probabilities assigned to any pair of observed values $(\tilde{\theta}^*_{j,i}, \tilde{\theta}^*_{l,i})$ is equal to $(\tilde{n}_{j,i}-\sigma)/(\tilde{n}_{l,i}-\sigma)$. If $\tilde{n}_{j,i} > \tilde{n}_{l,i}$, this is an increasing function of σ and, as σ increases the mass is reallocated from $\tilde{\theta}^*_{j,i}$ to $\tilde{\theta}^*_{l,i}$. This means that the sampling procedure tends to reinforce, among the observed clusters, those having higher frequencies. Such a reinforcement is analogous to the one discussed in exchangeable case in Lijoi, Mena and Prünster (2007a). In light of the above considerations the role of σ can then be summarized as follows: the larger σ the higher is the probability of generating a new value and at the same time the stronger is the reinforcement mechanism.

As far as the analysis of the dependence structure is concerned, it is useful to resort to the two extreme cases of total exchangeability (i.e. $z = 0$ implying $\tilde{p}_1 = \tilde{p}_2$ almost surely) and independence between $\tilde{p}_1$ and $\tilde{p}_2$ (implied by $z = 1$), since they provide useful hints for understanding the behaviour in intermediate situations corresponding to $z \in (0,1)$. When $z = 0$, then $\boldsymbol{\zeta}^{(1)} = \mathbf{0}_{n_1}$ and $\boldsymbol{\zeta}^{(2)} = \mathbf{0}_{n_2}$ (almost surely), where $\mathbf{0}_n$ is vector of 0s of size n. This is equivalent to considering the predictive distributions (13) and (14) with $\zeta_{n_\ell+1,\ell} = 0$, $\bar{n}_\ell = 0$, $\bar{n}_0 = n_1 + n_2$ and, therefore, $\bar{k}_\ell = 0$ and $\bar{k}_0 = k_1 + k_2$. In contrast, if $z = 1$ then $\boldsymbol{\zeta}^{(1)} = \mathbf{1}_{n_1}$ and $\boldsymbol{\zeta}^{(2)} = 2\,\mathbf{1}_{n_2}$ (almost surely), where $\mathbf{1}_n$ stands for a n–sized vector of 1s. This implies that $\zeta_{n_\ell+1,\ell} = \ell$, $\bar{n}_\ell = n_\ell$, $\bar{n}_0 = 0$ and, therefore, $\bar{k}_\ell = k_\ell$ and $\bar{k}_0 = 0$. In Figure 1 we compare the prior distribution of the total number of clusters $K^{(z)}_{n_1+n_2}$ for two vectors of observations of size $n_1 = n_2 = 200$, whose corresponding latent variables are governed by GM–$\mathscr{D}(c, z, P_0)$ and by GM–st(σ, z, P_0), where $z \in \{0,1\}$. The characterizing parameters c and σ are chosen so that $\mathbb{E}[K^{(0)}_{n_1+n_2}] \approx 5$ in 1(a), $\mathbb{E}[K^{(0)}_{n_1+n_2}] \approx 10$ in 1(b), $\mathbb{E}[K^{(0)}_{n_1+n_2}] \approx 20$ in 1(c) and $\mathbb{E}[K^{(0)}_{n_1+n_2}] \approx 50$ in 1(d). This choice reflects the idea of having two equivalent specifications in terms of the prior information on the number of clusters, thus making the comparison fair. The distribution of $K^{(0)}_{n_1+n_2}$ coincides with the distribution of the number of clusters for a single processes governing n_1+n_2 exchangeable latent variables; on the other extreme, the distribution of $K^{(1)}_{n_1+n_2}$ is the convolution of the distributions of

two random variables yielding the number of distinct values among n_1 and n_2 exchangeable random elements, respectively, governed by two independent processes.

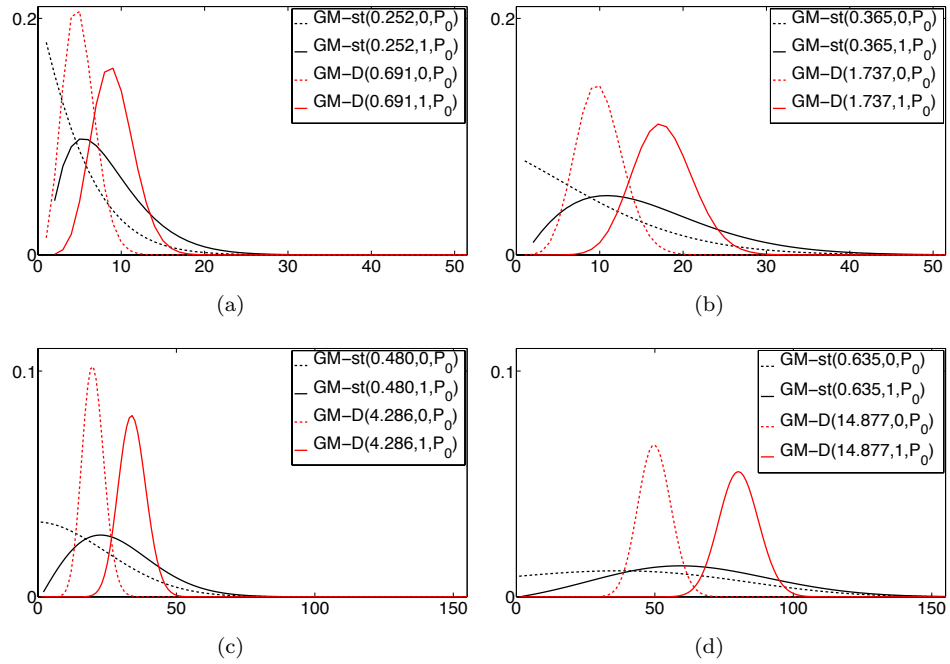


Figure 1: Prior distribution of the number of clusters of two samples of sizes $n_1 = n_2 = 200$ governed by GM–$\mathscr{D}$ processes (red lines) or GM–st processes (black lines) in the two extreme cases of $z = 0$ (dashed lines) and $z = 1$ (solid lines). The parameters c and σ are chosen so that, with $z = 0$, the expected number of clusters is approximately equal to 5 in (a), 10 in (b), 20 in (c) and 50 in (d).

It is apparent that both c and σ, for every $z \in \{0, 1\}$, have a role in determining the location: the larger they are, the greater is the expected number of clusters. This is in accordance with what was observed earlier, since the larger are c and σ, the greater is the probability of observing a new value in (13) and (14) respectively. Nonetheless, the model based on a GM–st process gives rise to much flatter distributions than those corresponding to the GM–$\mathscr{D}$ case. This is evident for both, small and large values of the parameters: in Figure 1(a), for example, GM–$\mathscr{D}$ models determine distributions that, although concentrated around small values, give very low probability

to the event $K^{(z)}_{n_1+n_2} \leq 2$ whereas, for GM–st models, the probability of the same event is remarkably larger and the distribution of $K^{(0)}_{n_1+n_2}$ has its mode in 1; from Figure 1(d) we see that, while the models based on GM–st processes determine flat distributions, for GM–$\mathscr{D}$ processes, the distribution of $K^{(z)}_{n_1+n_2}$ has a much smaller variance and is highly concentrated around its mode. Such findings are in accordance with the analysis developed, for the exchangeable case, in Lijoi, Mena and Prünster (2007a).

A further aspect to remark concerns the effect that, for each model, the dependence structure has on the clustering behavior. In this respect it should be recalled that, according to (13) and (14), the variable Z directly affects the probability of having a new value in the predictive distribution only for Dirichlet case. It is, then, expected that the degree of dependence, i.e. Z, is more influential in determining the clustering for the GM–$\mathscr{D}$ rather than for the GM–st process. This issue can be further investigated by comparing the two extreme cases, i.e. $z = 0$ vs $z = 1$: the closer they are in terms of the prior guess at the overall number of clusters in the two samples, the lower the influence of dependence on the clustering behaviour associated to the model. For this reason we consider the ratio $\mathbb{E}[K^{(1)}_{n_1+n_2}]/\mathbb{E}[K^{(0)}_{n_1+n_2}]$ and check for which model it is closer to 1. In particular, this analysis is performed by setting the values of the parameters c and σ such that $\mathbb{E}[K^{(0)}_{n_1+n_2}]$ takes on all possible values between 1 and $n_1 + n_2 = 400$. For each of these pairs (c, σ) the expected number of clusters $\mathbb{E}[K^{(1)}_{n_1+n_2}]$ ($z = 1$ meaning independence between $\tilde{p}_1$ and $\tilde{p}_2$) for both GM–$\mathscr{D}$ and GM–st models is computed so to yield the ratios depicted in Figure 2.

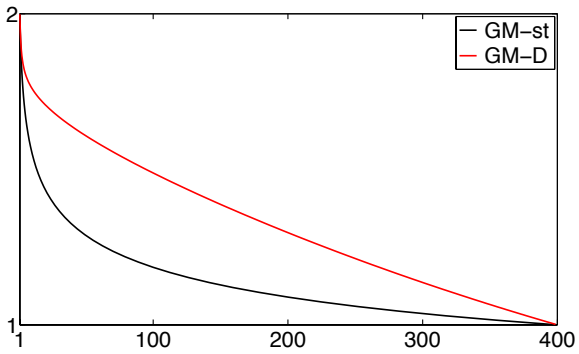


Figure 2: Plot of the ratio $\mathbb{E}[K^{(1)}_{n_1+n_2}]/\mathbb{E}[K^{(0)}_{n_1+n_2}]$, as a function of $\mathbb{E}[K^{(0)}_{n_1+n_2}]$, for mixture models based on GM–$\mathscr{D}$ and GM–st processes, with $n_1 = n_2 = 200$.

As expected, in both cases, the ratio is a decreasing function that tends to 2 when $\mathbb{E}[K^{(0)}_{n_1+n_2}]$ approaches 1, that is when $c \to 0$ and $\sigma \to 0$, and tends to 1 when $\mathbb{E}[K^{(0)}_{n_1+n_2}]$ approaches $n_1 + n_2$, that is when $c \to \infty$ and $\sigma \to 1$. More importantly, the curve is significantly lower for GM–st models than for GM–$\mathscr{D}$ models. This provides further evidence of the intuition according to which, in terms of expected number of clusters, the GM–st model is less sensitive to the specification of the parameter z. In light of the previous considerations it then comes to no surprise that these different predictive features are very influential in determining the clustering of the data. And, the insights gained on such predictive structures allow to understand the underlying reasons leading to the results obtained in the next section.

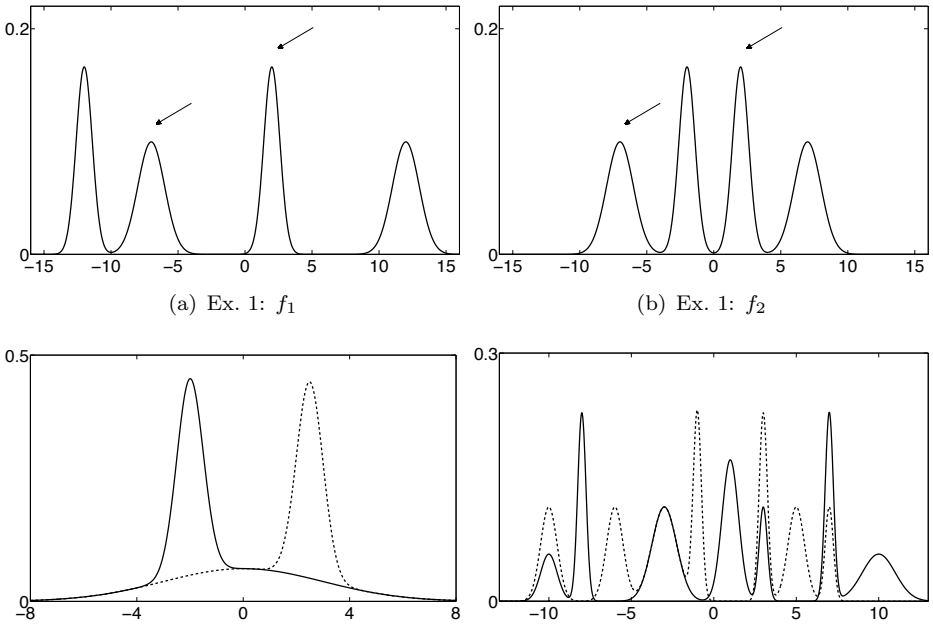
5 Simulation study

We perform an extensive simulation study for GM–dependent mixture models (10) by implementing the Gibbs sampling algorithm described in Section 3, and further detailed in the Appendix. We specifically focus on the posterior estimation of a pair of dependent densities and of the marginal clustering structures.

The datasets are generated from three different types of mixtures depicted in Figures 3. The first one is particularly simple and we use it to highlight specific features of the models. The other two examples refer to densities that are more challenging to estimate and include components that are not easily captured. They provide further support to the conclusions reached in the first example showing that GM–dependent mixtures can be successfully applied also in complex settings.

The goal of our analysis is two–fold. On the one hand, we aim at highlighting the borrowing strength phenomenon induced by the bivariate models if compared to the corresponding univariate analysis with the exchangeability; the latter corresponds to $(\tilde{p}_1, \tilde{p}_2)$ with Z being degenerate at $z = 1$. The second target is the comparison between the GM–$\mathscr{D}$ and the GM-st mixtures in terms of inference on the clustering structure of the data. The analysis of the performances of the two models in this numerical study essentially vouches the arguments that have emerged while investigating the predictive structure of the underlying processes.

Before entering the specific examples, we detail the models' specifications that are used henceforth. Unlike the previous sections, where Z was fixed



(a) Ex. 1: f_1 (b) Ex. 1: f_2

(c) Ex. 2: f_1 (cont. line) & f_2 (dashed line) (d) Ex. 3: f_1 (cont. line) & f_2 (dashed line)

Figure 3: [Examples 1,2 & 3]. True densities f_1 and f_2 generating the simulated datasets.

at a specific value z, here we introduce a prior distribution for it: this allows the data to provide information on the degree of dependence between $\tilde{p}_1$ and $\tilde{p}_2$. The specification is completed by an extension to the partially exchangeable case of the quite standard specification of Escobar and West (1995). In particular, we shall assume that $\theta = (M, V) \in \mathbb{R} \times \mathbb{R}^+$ and $h(\cdot\,; M, V)$ is a Gaussian density with mean M and variance V. We also take P_0 to be a normal/inverse–gamma distribution

$$P_0(\mathrm{d}M, \mathrm{d}V) = P_{0,1}(\mathrm{d}V)\, P_{0,2}(\mathrm{d}M \,|\, V),$$

with $P_{0,1}$ being an inverse–gamma probability distribution with parameters $(1, 1)$ and $P_{0,2}$ is Gaussian with mean m and variance τV. Moreover, the

corresponding hyperpriors are of the form

$$\begin{aligned} \tau^{-1} &\sim \text{Ga}(1/2, 50), \\ m &\sim \text{N}(\bar{D}, 2), \\ Z &\sim \text{U}(0,1), \\ c &\sim \text{Ga}(2,1) \qquad \text{in GM–}\mathscr{D}\text{ models}, \\ \sigma &\sim \text{U}(0,1) \qquad \text{in GM–st models}, \end{aligned} \tag{17}$$

where $\bar{D} = (\sum_{i=1}^{n_1} X_i + \sum_{j=1}^{n_2} Y_j)/(n_1 + n_2)$ is the over all sample mean. In the above specification, $\text{Ga}(a, b)$ stands for the gamma distribution with expected value a/b. From the specification in (17) one can immediately compute the *a priori* marginal expected number of components of the considered mixtures for different sample sizes. In the following examples we will consider sample sizes $n_\ell = 50$ and $n_\ell = 200$, which correspond to $\mathbb{E}[K_{50}] \cong 6.64$ and $\mathbb{E}[K_{200}] \cong 9.34$ for the GM–$\mathscr{D}$ mixtures and to $\mathbb{E}[K_{50}] \cong 13.32$ and $\mathbb{E}[K_{200}] \cong 39.67$ for GM–st mixtures. Henceforth, we shall slightly modify the notation and use $K_{\boldsymbol{X}}$ and $K_{\boldsymbol{Y}}$ in order to highlight the marginal number of clusters in the samples $\boldsymbol{X} = (X_1, \ldots, X_{n_1})$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_{n_2})$, respectively.

All estimates are based on 80000 iterations of the algorithm after 20000 burn–in sweeps.

5.1 Example 1

The data $\boldsymbol{X}$ and $\boldsymbol{Y}$ are generated as two independent samples of size $n_1 = n_2 = 200$, from densities f_1 and f_2, respectively, where

$$f_j = g_j + g_0, \qquad j = 1, 2,$$

with common component $g_0 \propto \text{N}(-7, 1) + \text{N}(2, 0.6)$ and idiosyncratic components $g_1 \propto \text{N}(-12, 0.6) + \text{N}(12, 1)$ and $g_2 \propto \text{N}(-2, 0.6) + \text{N}(7, 1)$. See Figure 3(a)–3(b). We then apply our algorithm to obtain estimates of: (i) the densities f_ℓ, with $\ell = 1, 2$ by means of the posterior expected value of the model; (ii) the number of clusters by means of the distribution of $K_{\boldsymbol{X}}$ and $K_{\boldsymbol{Y}}$.

First the density estimates for both GM–$\mathscr{D}(c, Z, P_0)$ and GM–st(σ, Z, P_0) mixtures are derived. These are reported in Figure 4 and show a good fit for both marginal densities and both models. It is also not surprising that the estimates obtained from the two models do not differ significantly: indeed, the differences at the latent level due to the mixing measures are smoothed

out by the density kernel and also one can always achieve a good fit with a larger number of components than actually needed.

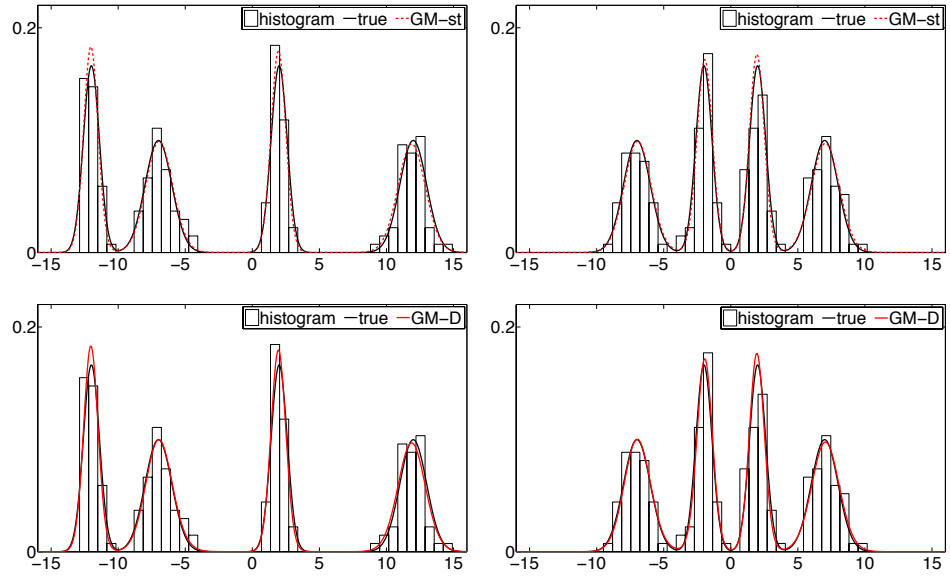
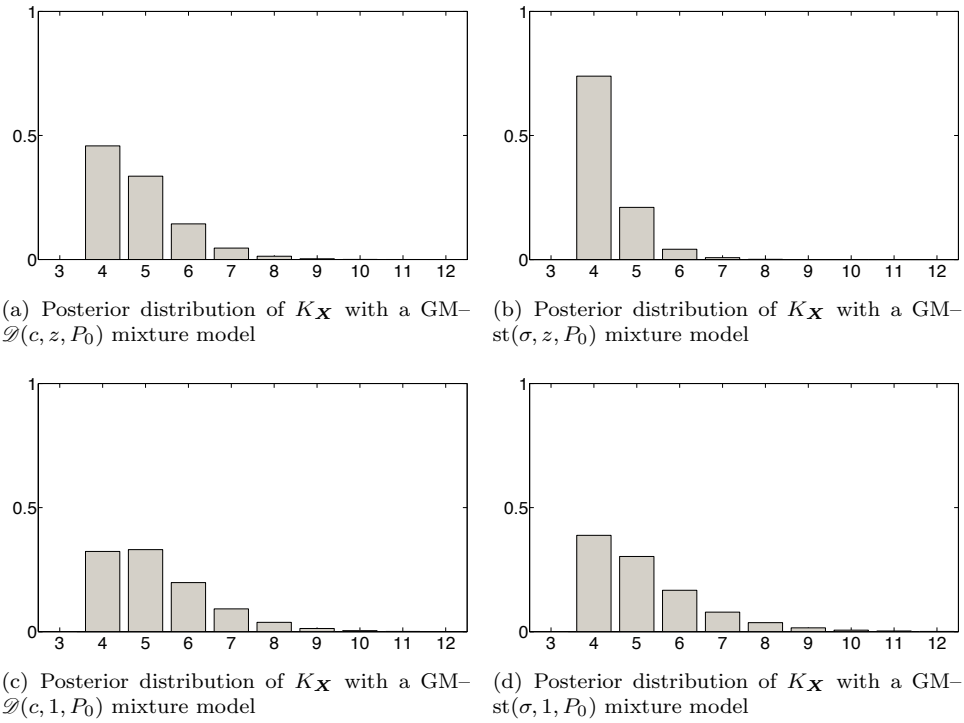


Figure 4: [Example 1]. True data generating densities (f_1 on the left column and f_2 on the right column) with histograms of the simulated data. Corresponding estimates are obtained with the GM–st(σ, Z, P_0) mixture model (first row) and the GM–$\mathscr{D}(c, Z, P_0)$ mixture model (second row).

It is to be noted that, since we rely on an MCMC procedure that incorporates the marginalization of the underlying RPMs, measures of variability of the density estimates are not part of the MCMC output. This is a well–known issue with marginal methods. A possible solution, though not straightforward to extend to the present context, is devised in Gelfand and Kottas (2002).

The most interesting aspects emerging from model comparison concern the analysis of the number of clusters. We first compare the dependent and the independent case, namely bivariate dependent and two univariate mixtures with GM–$\mathscr{D}$ and GM–st mixing measures. Figure 5 displays such a comparison and focuses on the first sample $\boldsymbol{X}$. Similar conclusions, even though not depicted here, hold true for $K_{\boldsymbol{Y}}$.



(a) Posterior distribution of $K_{\boldsymbol{X}}$ with a GM–$\mathscr{D}(c, z, P_0)$ mixture model

(b) Posterior distribution of $K_{\boldsymbol{X}}$ with a GM–st(σ, z, P_0) mixture model

(c) Posterior distribution of $K_{\boldsymbol{X}}$ with a GM–$\mathscr{D}(c, 1, P_0)$ mixture model

(d) Posterior distribution of $K_{\boldsymbol{X}}$ with a GM–st$(\sigma, 1, P_0)$ mixture model

Figure 5: [Example 1]. GM–$\mathscr{D}$ and GM–st mixtures (top row) vs. two independent univariate Dirichlet and normalized σ–stable process mixtures (bottom row): posterior distributions of the number of clusters for the first sample.

The superiority of the dependent partially exchangeable models is apparent: the marginal posterior distributions of the number of clusters $K_{\boldsymbol{X}}$, for both GM–$\mathscr{D}$ and GM–st models, tend to be more concentrated around the true value 4. Such differences are not surprising since they reveal that a phenomenon of borrowing strength applies when $Z < 1$, thus leading to a more reliable estimate of the number of clusters. The qualitative findings highlighted by Figure 5 are further corroborated by the numerical estimates in Table 1. Indeed, the estimates of $K_{\boldsymbol{X}}$ and $K_{\boldsymbol{Y}}$, with GM–$\mathscr{D}(c, Z, P_0)$ and GM–st(σ, Z, P_0) mixtures are closer to the true value than the estimates resulting from the GM–$\mathscr{D}(c, 1, P_0)$ and GM–st$(\sigma, 1, P_0)$ mixtures. This happens regardless as to whether the estimates are evaluated in terms of posterior expectations or in terms of maximum a posteriori values, $\hat{K}_{\boldsymbol{X}}$ and $\hat{K}_{\boldsymbol{Y}}$. More importantly, the posterior distributions of both $K_{\boldsymbol{X}}$ and $K_{\boldsymbol{Y}}$ obtained with GM–$\mathscr{D}(c, 1, P_0)$ and GM–st$(\sigma, 1, P_0)$ mixtures display a higher variabil-

Table 1: [Example 1]. GM–$\mathscr{D}$ mixture vs. independent univariate Dirichlet process mixtures (Rows 1 and 2) and GM–st mixture vs. independent univariate normalized σ–stable process mixtures (Rows 3 and 4): posterior expected number of clusters (Cols. 1 and 2), maximum a posteriori values ($\hat{K}_{\boldsymbol{X}}$, $\hat{K}_{\boldsymbol{Y}}$) and posterior probability of 6 or more clusters per sample (Cols. 5 and 6).

	$\mathbb{E}[K_{\boldsymbol{X}}\|\cdot]$	$\mathbb{E}[K_{\boldsymbol{Y}}\|\cdot]$	$\hat{K}_{\boldsymbol{X}}$	$\hat{K}_{\boldsymbol{Y}}$	$\mathbb{P}[K_{\boldsymbol{X}} \geq 6\|\cdot]$	$\mathbb{P}[K_{\boldsymbol{Y}} \geq 6\|\cdot]$
GM–$\mathscr{D}(c, Z, P_0)$	4.83	5.18	4	5	0.21	0.33
GM–$\mathscr{D}(c, 1, P_0)$	5.25	6.79	5	6	0.35	0.73
GM–st(σ, Z, P_0)	4.31	4.50	4	4	0.05	0.10
GM–st$(\sigma, 1, P_0)$	5.17	6.98	4	5	0.31	0.68

ity with heavier right–tails as can be ascertained by inspecting, e.g., the mass assigned to values of $K_{\boldsymbol{X}}$ and of $K_{\boldsymbol{Y}}$ greater than 6: such probability masses are significantly larger if compared to those yielded by GM–$\mathscr{D}(c, Z, P_0)$ and GM–st(σ, Z, P_0) mixtures. See also Figures 5(c)–5(d). Finally, note that, while the performances of GM–$\mathscr{D}(c, 1, P_0)$ and GM–st$(\sigma, 1, P_0)$ are rather similar with the latter showing a slightly larger concentration around the true value 4, the GM–st(σ, Z, P_0) mixture seems to be superior to the GM–$\mathscr{D}(c, Z, P_0)$ in detecting the correct number of components, as can be appreciated through both Figures 5(a)–5(b) and Table 1.

5.1.1 Fixing the parameters σ and c

The previous analysis clearly shows the superiority of dependent models over models with independent $\tilde{p}_1$ and $\tilde{p}_2$. However, in order to make a fair comparison between GM–$\mathscr{D}(c, Z, P_0)$ and GM–st(σ, Z, P_0) mixtures, one needs to adopt "similar" prior specifications for the vectors (c, Z) and (σ, Z). Hence, we assume the same prior for Z in both mixtures and set degenerate priors for σ and c in (17). Given we want to compare their performance in terms of clustering, c and σ are fixed so to obtain for both mixtures the same marginal *a priori* expected number of clusters. Hence, we shall consider values of σ and c such that under both a Dirichlet and a normalized σ–stable process one has

$$\mathbb{E}[K_{\boldsymbol{X}}] = \mathbb{E}[K_{\boldsymbol{Y}}] \cong 15. \tag{18}$$

Solving (18) leads to set $c = c_1 = 3.587$ and $\sigma = \sigma_1 = 0.4884$, respectively. The idea is to specify parameter values yielding a prior opinion that is far from the truth (i.e. 4 clusters) and identify which mixture model better detects, on the basis of the information conveyed by the data, the correct number of clusters. Alternatively, the parameters c and σ can be fixed in such a way that they yield similar prior variance structures for $\tilde{p}_1$ and $\tilde{p}_2$ in both models. To this end, one can resort to (James, Lijoi and Prünster, 2006, Proposition 1) which implies that $\text{Var}[\tilde{p}_\ell(B)]$ is the same both for a Dirichlet and a normalized σ–stable process, for any B in $\mathscr{X}$ and $\ell = 1, 2$, if and only if $(c+1)^{-1} = 1 - \sigma$. Hence, for $c = c_1$ the variances match if $\sigma = \sigma_2 = 0.782$. On the other hand, if $\sigma = \sigma_1$ the variances coincide if $c = c_2 = 0.9547$. This leads to draw three comparisons of mixtures: (i) GM–$\mathscr{D}(c_1, Z, P_0)$ vs GM–st(σ_1, Z, P_0); (ii) GM–$\mathscr{D}(c_1, Z, P_0)$ vs GM–st(σ_2, Z, P_0); (iii) GM–$\mathscr{D}(c_2, Z, P_0)$ vs GM–st(σ_1, Z, P_0). Posterior inferences on the marginal number of clusters, $K_{\boldsymbol{X}}$ and $K_{\boldsymbol{Y}}$, are summarized in Table 2.

Table 2: [Example 1]. Comparisons between GM–$\mathscr{D}(c, Z, P_0)$ vs. GM–st(σ, Z, P_0) mixtures. The parameters c and σ are fixed in such a way that: (i) condition (18) holds true (Row 1 vs. Row 3); (ii) marginal prior variance structures of $\tilde{p}_i$ in both models match (Row 1 vs. Row 2 & Row 3 vs. Row 4).

	$\mathbb{E}[K]$	$\mathbb{E}[K_{\boldsymbol{X}}\|\cdot]$	$\mathbb{E}[K_{\boldsymbol{Y}}\|\cdot]$	$\hat{K}_{\boldsymbol{X}}$	$\hat{K}_{\boldsymbol{Y}}$	$\mathbb{P}[K_{\boldsymbol{X}} \geq 6\|\cdot]$	$\mathbb{P}[K_{\boldsymbol{Y}} \geq 6\|\cdot]$
GM-$\mathscr{D}(c_1, Z, P_0)$	15	6.96	7.82	7	7	0.81	0.91
GM–st(σ_2, Z, P_0)	67.95	4.81	5.27	4	4	0.20	0.35
GM–st(σ_1, Z, P_0)	15	4.62	4.96	4	4	0.14	0.25
GM-$\mathscr{D}(c_2, Z, P_0)$	5.69	4.81	5.12	4	5	0.19	0.31

These unequivocally show a better performance of the GM–st mixture. For example, the posterior distributions of the number of clusters for the GM–$\mathscr{D}(c_1, Z, P_0)$ mixture feature much heavier right–tails if compared to the estimates resulting from the GM–st(σ_1, Z, P_0) mixture. This means that the GM–st mixture yields estimates of the posterior distributions of both $K_{\boldsymbol{X}}$ and $K_{\boldsymbol{Y}}$ concentrated around the correct number of components of the mixtures that have actually generated the data, despite the prior misspecification. Also in terms of the posterior estimates of $K_{\boldsymbol{X}}$ and $K_{\boldsymbol{Y}}$, reported in Table 2, the GM–st mixture stands out regardless of whether they are estimated by means of posterior expectation or by maximum *a posteriori*

values. Similar conclusions can be drawn for the other two comparisons where the prior marginal variance structure is taken to be the same under both models. Of these, the most interesting is, indeed, the case where the GM–st(σ_2, Z, P_0) mixture outperforms the GM(c_1, Z, P_0) mixture: though the GM–st mixture specification yields an expected number of components much larger than that of the GM–$\mathscr{D}$ mixture, it is still able to recover the correct number of components. This can also be noted for the last comparison, summarized in the third and fourth lines of Table 2, where in the GM–$\mathscr{D}$ case the prior for both $K_{\boldsymbol{X}}$ and $K_{\boldsymbol{Y}}$ is concentrated around the true number of components, in contrast to the GM–st process for which the prior expected number of components is 15. Despite this prior misspecification, the GM–st(σ_1, z, P_0) mixture leads to posterior estimates of $K_{\boldsymbol{X}}$ and $K_{\boldsymbol{Y}}$ close to the truth and posterior distributions for $K_{\boldsymbol{X}}$ and $K_{\boldsymbol{Y}}$ with lighter right tails.

The empirical evidence displayed in Table 2 shows that GM–st models are less sensitive to the misspecification of their characterizing parameters. This could be explained by the fact that such models give rise to prior distributions for the number of clusters that are much flatter, i.e. less informative, than the corresponding distributions for GM–$\mathscr{D}$ mixtures. See also Figure 1.

5.1.2 The role of σ in GM–st mixtures

Since the posterior results obtained so far suggest that GM–st mixture models typically lead to a more effective detection of the clustering structure of the observations, it is worth analyzing empirically the role of $\sigma \in (0, 1)$ in such models. To this end we fix a grid of values, $\{0.1, 0.3, 0.5, 0.7, 0.9\}$, for σ and determine the corresponding posterior estimates of the densities f_1 and f_2 and the posterior distributions of the number of clusters $K_{\boldsymbol{X}}$ and $K_{\boldsymbol{Y}}$ in the two samples. Two interesting indications on the role of σ can be deduced from this analysis. First, the density estimates, although not reported, are not significantly sensitive to the choice of σ. Second, as shown by Table 3, the value of σ affects the distribution of the number of clusters.

Small values of σ correspond to centering the prior of the number of components on small values: therefore, it is natural that in this specific case, where we have a small number of components (i.e. 4), a small value of σ provides a better fit. Nonetheless, as already seen before we performed the variance match, the model shows to adapt reasonably well even for large values of σ, which in this example correspond to severe prior misspecification. Such prior specification issues could be circumvented by placing a prior on σ: as

Table 3: [Example 1]. GM–st(σ, Z, P_0) mixture: estimated number of clusters and posterior probability of 6 or more clusters per sample for different values of σ.

σ	0.1	0.3	0.5	0.7	0.9	random
$\mathbb{E}[K_{\boldsymbol{X}}\|\cdot]$	4.17	4.44	4.65	4.74	4.83	4.31
$\mathbb{E}[K_{\boldsymbol{Y}}\|\cdot]$	4.24	4.66	4.96	5.18	5.32	4.50
$\mathbb{P}[K_{\boldsymbol{X}} \geq 6\|\cdot]$	0.04	0.08	0.15	0.18	0.21	0.05
$\mathbb{P}[K_{\boldsymbol{Y}} \geq 6\|\cdot]$	0.05	0.15	0.25	0.33	0.36	0.10

mentioned before, in (17) we have set a uniform prior on $(0, 1)$ and the corresponding posterior estimate for σ is equal to 0.21. But it is evident that GM–st mixtures work fairly well also for fixed (and possibly misspecified) σ. In contrast, for GM–$\mathscr{D}$ mixtures putting a prior on the parameter c is crucial in order to reasonably detect the clustering structure but, depending on the specific clustering structure, one may even end up with the unpleasant feature of the inferences depending on the type of prior specified for c. See Dorazio et al. (2008).

5.2 Example 2

Example 1 has served as a "toy example" useful for displaying specific features of GM–$\mathscr{D}$ and GM–st mixtures and for drawing a comparison between the two. Here, we consider a more challenging situation. A first source of difficulty is due to the sizes of the two independent samples. We consider an unbalanced situation where the size of the first sample, $n_1 = 200$, is much larger than the size of the second sample, $n_2 = 50$. This is combined with the choice of data generating mixtures that have both two components which are very close one to the other. More precisely we consider two densities f_1 and f_2 defined as

$$f_1 \sim \frac{1}{2}\mathrm{N}(0, 3) + \frac{1}{2}\mathrm{N}(-2, 0.5), \qquad f_2 \sim \frac{1}{2}\mathrm{N}(0, 3) + \frac{1}{2}\mathrm{N}(2, 0.5), \tag{19}$$

which share one component and are depicted in Figure 3(c). In order to estimate f_1, f_2, $K_{\boldsymbol{X}}$ and $K_{\boldsymbol{Y}}$ we rely, as before, on the prior specifications (17). The density estimates, not displayed here, again show a good fit. In terms of inference on the clustering structure, the dependent model heavily benefits from the borrowing strength effect as apparent from Table 4, where the results are compared to the independent case ($Z = 1$). Even for the second

(under-represented) sample, both mixtures are able to get close to the correct number of components. More importantly, the posterior distributions of $K_{\boldsymbol{X}}$ and $K_{\boldsymbol{Y}}$ in the dependent case are much more concentrated around the true number of components assigning significantly less probability to 4 or more components.

Table 4: [Example 2]. GM–$\mathscr{D}$ mixture vs. independent univariate Dirichlet process mixtures (Rows 1 and 2) and GM–st mixture vs. independent univariate normalized σ–stable process mixtures (Rows 3 and 4): posterior expected number of clusters (Cols. 1 and 2), maximum a posteriori values ($\hat{K}_{\boldsymbol{X}}$, $\hat{K}_{\boldsymbol{Y}}$) and posterior probability of 4 or more clusters per sample (Cols. 5 and 6).

	$\mathbb{E}[K_{\boldsymbol{X}}\|\cdot]$	$\mathbb{E}[K_{\boldsymbol{Y}}\|\cdot]$	$\hat{K}_{\boldsymbol{X}}$	$\hat{K}_{\boldsymbol{Y}}$	$\mathbb{P}[K_{\boldsymbol{X}} \geq 4\|\cdot]$	$\mathbb{P}[K_{\boldsymbol{Y}} \geq 4\|\cdot]$
GM–$\mathscr{D}(c, Z, P_0)$	2.76	3.17	2	2	0.19	0.33
GM–$\mathscr{D}(c, 1, P_0)$	3.44	4.06	2	3	0.38	0.53
GM–st(σ, Z, P_0)	2.22	2.23	2	2	0.03	0.03
GM–st$(\sigma, 1, P_0)$	3.23	2.86	2	2	0.27	0.22

As to the comparison of GM–$\mathscr{D}$ and GM–st mixtures, both yield roughly the same estimates for the marginal number of components, but the GM–st(σ, Z, P_0) mixture is by far superior when looking at the variability of the posterior distributions of $K_{\boldsymbol{X}}$ and $K_{\boldsymbol{Y}}$. This confirms the findings of Section 5.1: GM–st models outperform GM–$\mathscr{D}$ models when it comes to drawing inferences on the clustering structure featured by the data. Such conclusions are even more apparent if an analysis along the lines of Section 5.1.1, and not reported here, is carried out.

5.3 Example 3

In this final example we consider data generated by mixtures with a large number of modes and with components having different weights. We generate two independent samples of size $n_1 = n_2 = 200$ from densities f_1 and f_2 defined as mixtures of seven normals with four of them in common. Moreover, the common mixed densities are weighted differently in the two mixtures. More precisely, we set

$$f_1 \sim \sum_{i=1}^{10} a_i \mathrm{N}(\mu_i, \sigma_i), \qquad f_2 \sim \sum_{i=1}^{10} b_i \mathrm{N}(\mu_i, \sigma_i), \tag{20}$$

where the vectors of means and standard deviations are respectively equal to $\boldsymbol{\mu} = (-10, -8, -6, -3, -1, 1, 3, 5, 7, 10)$ and $\boldsymbol{\sigma} = \frac{1}{4}(2, 1, 2, 3, 1, 2, 1, 2, 1, 4)$. The weights in (20) are identified by $\boldsymbol{a} = (1, 2, 0, 3, 0, 3, 1, 0, 2, 2)/14$ and $\boldsymbol{b} = (2, 0, 2, 3, 2, 0, 2, 2, 1, 0)/14$. The two mixtures are displayed in Figure 3(d) which allows to visualize the different weights assigned to the components shared by the mixtures defining f_1 and f_2. Density estimates are represented in Figure 6 and even in such a challenging example we achieve a satisfactory fit. Moreover, as expected, they do not significantly differ among the two GM–$\mathscr{D}$ and GM–st models.

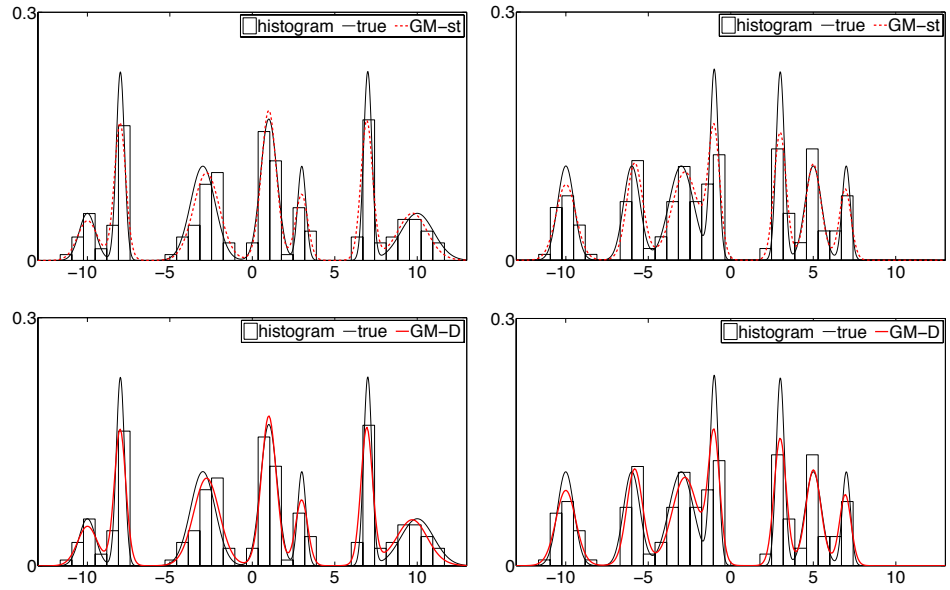


Figure 6: [Example 3]. True data generating densities (f_1 on the left column and f_2 on the right column) with histograms of the simulated data. Corresponding estimates are obtained with the GM–st(σ, Z, P_0) mixture model (first row) and the GM–$\mathscr{D}(c, Z, P_0)$ mixture model (second row).

The posterior inferences concerning the marginal clustering structure are reported in Table 5 and lead to results similar to those obtained in Sections 5.1 and 5.2. Indeed, posterior estimates of $K_{\boldsymbol{X}}$ and of $K_{\boldsymbol{Y}}$ based on both GM–$\mathscr{D}(c, Z, P_0)$ and GM–st(σ, Z, P_0) mixtures are very close to the actual value that has generated the data. However, the lighter right–tails

Table 5: [Example 3]. GM–$\mathscr{D}$ mixture vs. independent univariate Dirichlet process mixtures (Rows 1 and 2) and GM–st mixture vs. independent univariate normalized σ–stable process mixtures (Rows 3 and 4): posterior expected number of clusters (Cols. 1 and 2), maximum a posteriori values ($\hat{K}_{\boldsymbol{X}}$, $\hat{K}_{\boldsymbol{Y}}$) and posterior probability of 9 or more clusters per sample (Cols. 5 and 6).

	$\mathbb{E}[K_{\boldsymbol{X}}\|\cdot]$	$\mathbb{E}[K_{\boldsymbol{Y}}\|\cdot]$	$\hat{K}_{\boldsymbol{X}}$	$\hat{K}_{\boldsymbol{Y}}$	$\mathbb{P}[K_{\boldsymbol{X}} \geq 9\|\cdot]$	$\mathbb{P}[K_{\boldsymbol{Y}} \geq 9\|\cdot]$
GM–$\mathscr{D}(c, Z, P_0)$	8.03	7.96	7	7	0.27	0.25
GM–$\mathscr{D}(c, 1, P_0)$	9.49	8.95	9	8	0.64	0.52
GM–st(σ, Z, P_0)	7.45	7.39	7	7	0.08	0.06
GM–st$(\sigma, 1, P_0)$	9.93	9.27	9	8	0.60	0.54

of the GM–st mixture that are highlighted in the last two columns of Table 5 suggest that the dependent GM–st model is preferable. Moreover, a comparison with univariate mixtures ($Z = 1$) shows again the beneficial effect of the borrowing strength phenomenon. In this respect, estimates arising from both the GM–$\mathscr{D}(c, 1, P_0)$ and GM–st$(\sigma, 1, P_0)$ processes are farther away from the true number of components if compared to those obtained through the corresponding bivariate dependent model. Furthermore, in the case $Z = 1$ the posterior probability assigned to number of clusters larger than 9 is significantly larger as can be seen from the last two columns of Table 5.

6 Concluding remarks

Both GM–$\mathscr{D}$ and GM–st dependent mixture models exhibit a good performance in terms of density estimation even in challenging problems. However, when it comes to estimation of the number of clusters they show significantly different features. First, GM–st models stand out for being less sensitive to the specification of their characterizing parameters. This property can be explained by the flatness of the prior distribution for the number of clusters induced by such models and is a result of the predictive structure thoroughly described in Section 4. Second, while for both classes of mixture models, the borrowing strength is remarkable, it seems that GM–st models better profit from the borrowed information by greatly improving their performance. Overall, there is clear evidence that GM–st models feature an improved capability of learning from the data and detecting the correct number of clusters. Such a phenomenon, that was first noted in the

univariate case in Lijoi, Mena and Prünster (2007a), is confirmed in the dependent case. Moreover, we showed that it also has a positive influence on the borrowing information. These findings show that models based on dependent normalized σ–stable processes represent an appealing alternative to the commonly used models based on dependent DPs.

Acknowledgments

A. Lijoi and I. Prünster are supported by the European Research Council (ERC) through StG "N-BNP" 306406. A. Lijoi and I. Prünster gratefully acknowledge the hospitality of ICERM at Brown University, where important parts of this work were carried out during the Research Program "Computational Challenges in Probability".

References

Cifarelli, D.M. and Regazzini, E. (1978). Problemi statistici non parametrici in condizioni di scambiabilità parziale. *Quaderni Istituto Matematica Finanziaria, Università di Torino* Serie III, **12**. English translation available at: http://www.unibocconi.it/wps/allegatiCTP/CR-Scambparz[1].20080528.135739.pdf

Daley, D.J. and Vere–Jones, D. (1988). *An introduction to the theory of point processes.* Springer, New York.

Dorazio, R.M., Mukherjee, B., Zhang, L., Ghosh, M., Jelks, H.L. and Jordan, F. (2008). Modeling unobserved sources of heterogeneity in animal abundance using a Dirichlet process prior. *Biometrics* **64**, 635–644.

Escobar, M.D. and West, M. (1995). Bayesian density estimation and inference using mixtures, *J. Amer. Statist. Assoc.* **90**, 577–588.

de Finetti, B. (1938). Sur la condition d'equivalence partielle. *Actualités scientifiques et industrielles* **739**, 5–18. Herman, Paris.

Gelfand, A.E. and Kottas, A. (2002). A computational approach for full nonparametric Bayesian inference under Dirichlet process mixture models. *J. Comp. Gaph. Statist.* **11**, 289–305.

Griffiths, R.C. and Milne, R.K. (1978). A class of bivariate Poisson processes. *J. Mult. Anal.* **8**, 380–395.

Hartigan, J.A. (1990). Partition models. *Comm. Statist. Theory Methods* **19**, 2745–2756.

Hatjispyros, S.J., Nicoleris, T. and Walker, S.G. (2011). Dependent mixtures of Dirichlet processes. *Comput. Statist. Data Anal.* **55**, 2011–2025.

Hjort, N.L., Holmes, C.C. Müller, P. and Walker, S.G. (Eds.) (2010). *Bayesian Nonparametrics*. Cambridge University Press, Cambridge.

Ishwaran, H. and James, L.F. (2001). Gibbs sampling methods for stick-breaking priors. *J. Amer. Statist. Assoc.* **96**, 161–173.

Ishwaran, H. and James, L.F. (2003). Generalized weighted Chinese restaurant processes for species sampling mixture models. *Statist. Sinica* **13**, 1211–1235.

James, L.F., Lijoi, A. and Prünster, I. (2006). Conjugacy as a distinctive feature of the Dirichlet process. *Scand. J. Statist. 33*, 105–120.

James, L.F., Lijoi, A. and Prünster, I. (2009). Posterior analysis for normalized random measures with independent increments. *Scand. J. Stat.* **36**, 76–97.

Jara, A., Hanson, T., Quintana, F.A., Müller, P. and Rosner, G. (2011) DPpackage: Bayesian Non- and Semi-parametric Modelling in R. *Journal of Statistical Software* **40**, 1–30.

Leon–Novelo, L., Nebiyou Bekele, B., Müller, P., Quintana, F.A. and Wathen, K. (2012). Borrowing strength with non-exchangeable priors over subpopulations. *Biometrics* **68**, 550–558.

Lijoi, A., Mena, R.H. and Prünster, I. (2005). Hierarchical mixture modelling with normalized inverse Gaussian priors. *J. Amer. Stat. Assoc.* **100**, 1278–1291.

Lijoi, A., Mena, R.H. and Prünster, I. (2007a). Controlling the reinforcement in Bayesian non-parametric mixture models. *J. R. Stat. Soc. Ser. B* **69**, 715–740.

Lijoi, A., Mena, R.H. and Prünster, I. (2007b). Bayesian nonparametric estimation of the probability of discovering new species. *Biometrika* **94**, 769–786.

Lijoi, A., Nipoti, B. and Prünster, I. (2013). Bayesian inference via dependent normalized completely random measures. *Bernoulli*, DOI: 10.3150/13-BEJ521.

Lijoi, A. and Prünster, I. (2010). Models beyond the Dirichlet process. In *Bayesian Nonparametrics* (Hjort, N.L., Holmes, C.C. Müller, P., Walker, S.G. Eds.), pp. 80–136. Cambridge University Press, Cambridge.

Lo, A. (1984). On a class of Bayesian nonparametric estimates. I. Density estimates. *Ann. Statist.* **12**, 351–357.

MacEachern, S.N. (1999). Dependent nonparametric processes. In *ASA Proceedings of the Section on Bayesian Statistical Science*, Alexandria, VA: American Statistical Association.

MacEachern, S.N. (2000). Dependent Dirichlet processes. *Technical Report*, Ohio State University.

Müller, P., Quintana, F.A. and Rosner, G. (2004). A method for combining inference across related nonparametric Bayesian models. *J. Roy. Statist. Soc. Ser. B* **66**, 735–749.

Müller, P., Quintana, F.A. and Rosner, G. L. (2011). A product partition model with regression on covariates. *J. Comput. Graph. Statist.* **20**, 260–278.

Nipoti, B. (2011). *Dependent completely random measures and statistical applications.* Ph.D. thesis. Department of Mathematics, University of Pavia.

Petrone, S. and Raftery A.E. (1997). A note on the Dirichlet process prior in Bayesian nonparametric inference with partial exchangeability. *Statist. Probab. Lett.* **36**, 69–83

Regazzini, E., Lijoi, A. and Prünster, I. (2003). Distributional results for means of random measures with independent increments. *Ann. Statist.* **31**, 560–585.

Appendix. Full conditional distributions

Here we provide the full conditional distributions for random variables and hyperparameters involved in the model introduced in (10) and (12), with the hyperpriors specified in (17). For a more general and detailed treatment, refer to Lijoi, Nipoti and Prünster (2013).

For each $\ell = 1, 2$, we call $\boldsymbol{\zeta}_*^{(\ell)}$ the vector of the auxiliary random variables corresponding to the distinct values of $\boldsymbol{\theta}^{(\ell)}$ and we let $\boldsymbol{\zeta}_{-j,*}^{(\ell)}$ denote the vector obtained by removing the j–th component from $\boldsymbol{\zeta}_*^{(\ell)}$. Moreover, we indicate with $\boldsymbol{h}$ the vector of all the hyperparameters involved in the model, that is z, c, τ and m for GM–$\mathscr{D}$ model and z, σ, τ and m for GM–st model . We introduce the notation

$$\pi_{j,1}(x) := \mathbb{P}[\zeta_{j,1}^* = x \,|\, \boldsymbol{\zeta}_{-j,*}^{(1)}, \boldsymbol{\zeta}_*^{(2)}, \boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{h}].$$

If $\theta_{j,1}^*$ does not coincide with any of the distinct values of the latent variables for the second sample, then, for the GM–$\mathscr{D}$ model, we have

$$\pi_{j,1}(x) \propto \mathbb{1}_{\{0,1\}}(x)\, \frac{z^x\,(1-z)^{1-x}}{(\alpha)_{n_2}(\beta_x)_{n_2}} \times {}_3F_2\left(\alpha - cz + n_1 - \bar{n}_{-j,1} - x n_{j,1}^*, n_1, n_2; \alpha + n_1, \beta_x + n_2; 1\right), \quad (21)$$

where $(a)_n$ is the Pochhammer symbol, ${}_3F_2$ denotes the generalized hypergeometric function, $\bar{n}_{-j,1} := \sum_{i \neq j} n^*_{i,1}\zeta^*_{i,1}$ and α and β_x are defined as $\alpha = c + n_2 - \bar{n}_2$ and $\beta_x = c + n_1 - \bar{n}_{-j,1} - x n^*_{j,1}$. For the GM–st model, we have

$$\pi_{j,1}(x) \propto \mathbb{1}_{\{0,1\}}(x)\, z^x (1-z)^{1-x} \times \int_0^1 \frac{w^{n-\bar{n}_{-j,1}-xn^*_{j,1}+(\bar{k}_{-j,1}+x)\sigma-1}(1-w)^{n_2-\bar{n}_2+\bar{k}_2\sigma-1}}{\{1-z+zw^\sigma+z(1-w)^\sigma\}^k}\,\mathrm{d}w, \quad (22)$$

where $\bar{k}_{-j,1} = \sum_{i\neq j}\zeta^*_{i,1}$. For both (21) and (22), the normalizing constant is determined by $\pi_{j,i}(0) + \pi_{j,i}(1) = 1$. The full conditionals for the $\zeta^*_{j,2}$ can be determined analogously.

As for the latent variables $\boldsymbol{\theta}^{(1)}$ and $\boldsymbol{\theta}^{(2)}$, one can sample $\theta_{j,\ell}$ from

$$w_0 P^*_{j,\ell}(\mathrm{d}\theta) + \sum_{i \in \mathcal{J}_{-j,\zeta_{j,\ell}}} w_i \delta_{\tilde{\theta}^*_{i,\zeta_{j,\ell}}}(\mathrm{d}\theta), \quad (23)$$

where $\mathcal{J}_{-j,\zeta_{j,\ell}}$ is the set of indices of distinct values from the urn labeled $\zeta_{j,\ell}$ after excluding $\theta_{j,\ell}$. For the GM–$\mathscr{D}$ model, the weights in (23) are given by

$$\begin{aligned} w_0 &\propto c\,\frac{\bar{z}\,(1+\tau)}{[2(1+\tau)+(m-x_{j,\ell})^2]^{3/2}}, \\ w_i &\propto n^{(-j)}_{i,\ell}\,\frac{1}{\sqrt{2\pi\tilde{V}^*_{i,\zeta_{j,\ell}}}}\exp\left\{-\frac{(x_{j,\ell}-\tilde{M}^*_{i,\zeta_{j,\ell}})^2}{2\tilde{V}^*_{i,\zeta_{j,\ell}}}\right\}, \end{aligned} \quad (24)$$

where $\bar{z} = (1-z)\mathbb{1}_{\{0\}}(\zeta_{j,\ell}) + z\mathbb{1}_{\{\ell\}}(\zeta_{j,\ell})$, $n^{(-j)}_{i,\ell}$ is the size of the cluster containing $\tilde{\theta}^*_{i,\zeta_{j,\ell}}$, after deleting $\theta_{j,\ell}$, and $\tilde{\theta}^*_{i,\zeta_{j,\ell}} = (\tilde{M}^*_{i,\zeta_{j,\ell}}, \tilde{V}^*_{i,\zeta_{j,\ell}})$.

For the GM–st model, the weights in (23) are given by

$$\begin{aligned} w_0 &\propto k_{-j,\zeta_{j,1}}\sigma\,\frac{\bar{z}\,(1+\tau)}{[2(1+\tau)+(m-x_{j,\ell})^2]^{3/2}}, \\ w_i &\propto \left(n^{(-j)}_{i,\ell}-\sigma\right)\,\frac{1}{\sqrt{2\pi\tilde{V}^*_{i,\zeta_{j,\ell}}}}\exp\left\{-\frac{(x_{j,\ell}-\tilde{M}^*_{i,\zeta_{j,\ell}})^2}{2\tilde{V}^*_{i,\zeta_{j,\ell}}}\right\}. \end{aligned} \quad (25)$$

The distribution $P^*_{j,\ell}$ of a new value $\theta = (M, V)$ in (23) is again a normal/inverse–gamma distribution, that is

$$P^*_{j,\ell}(\mathrm{d}M, \mathrm{d}V) = P^{(1)}_{j,\ell}(\mathrm{d}V)\, P^{(2)}_{j,\ell}(\mathrm{d}M \,|\, V), \quad (26)$$

with $P_{0,1}^{(1)}$ being an inverse–gamma probability distribution with parameters $(3/2, 1+(m-x_{j,\ell})^2/(2(\tau+1)))$ and $P_{0,2}^{(2)}$ Gaussian with mean $(\tau x_{j,\ell}+m)/(\tau+1)$ and variance $\tau/(\tau+1)V$.

Let now $\boldsymbol{D}_{-r}$ stand for the set of all random variables of the model (that is, latent variables $(\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)})$, auxiliary variables $(\boldsymbol{\zeta}^{(1)}, \boldsymbol{\zeta}^{(2)})$ and hyperparameters $\boldsymbol{h}$) but r. As for the full conditional for z, one has, for the GM-$\mathscr{D}$ model,

$$\kappa_z(z \,|\, \boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{D}_{-z}) \propto \mathbb{1}_{(0,1)}(z)\, z^{\bar{k}_1+\bar{k}_2}(1-z)^{\bar{k}_0} \times {}_3F_2\left(\alpha - cz + n_1 - \bar{n}_1, n_1, n_2; \alpha + n_1, \beta + n_2; 1\right),$$

where $\alpha = c + n_2 - \bar{n}_2$ and $\beta = c + n_1 - \bar{n}_1$. For the GM-st model, one has

$$\kappa_z(z \,|\, \boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{D}_{-z}) \propto \mathbb{1}_{(0,1)}(z)\, z^{\bar{k}_1+\bar{k}_2}(1-z)^{\bar{k}_0} \times \int_0^1 \frac{w^{n_1-\bar{n}_1+\bar{k}_1\sigma-1}(1-w)^{n_2-\bar{n}_2+\bar{k}_2\sigma-1}}{\{1-z+zw^\sigma+z(1-w)^\sigma\}^k}\mathrm{d}w. \quad (27)$$

As for the parameters c and σ that characterize respectively GM–$\mathscr{D}$ and GM-st models, we have

$$\kappa_c(c \,|\, \boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{D}_{-c}) \propto \frac{c^{k+1}\mathrm{e}^{-c}}{(\alpha)_{n_1}(\beta)_{n_2}} \times {}_3F_2\left(\alpha - cz + n_1 - \bar{n}_1, n_1, n_2; \alpha + n_1, \beta + n_2; 1\right),$$

and

$$\kappa_\sigma(\sigma \,|\, \boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{D}_{-\sigma}) \propto \mathbb{1}_{(0,1)}(\sigma)\, \sigma^{k-1}\, \xi_\sigma(\boldsymbol{n}^{(1)}, \boldsymbol{n}^{(2)}, \boldsymbol{n}^{(0)}) \times \int_0^1 \frac{w^{n_1-\bar{n}_1+\bar{k}_1\sigma-1}(1-w)^{n_2-\bar{n}_2+\bar{k}_2\sigma-1}}{\{1-z+zw^\sigma+z(1-w)^\sigma\}^k}\mathrm{d}w, \quad (28)$$

where

$$\xi_\sigma(\boldsymbol{n}^{(1)}, \boldsymbol{n}^{(2)}, \boldsymbol{n}^{(0)}) = \prod_{j=1}^{k_1}(1-\sigma)_{n^*_{j,1}-1} \prod_{i=1}^{k_2}(1-\sigma)_{n^*_{i,2}-1} \prod_{r=1}^{m}(1-\sigma)_{n^*_{r,0}-1}.$$

Notice that a numerical evaluation of the integrals in (27) and (28) is straightforward.

Finally, for both GM–$\mathscr{D}$ and GM–st models, τ and m are sampled from the following distributions

$$\tau \,|\, (\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{D}_{-\tau}) \sim \mathrm{IG}\left(\frac{1+k_0+k_1+k_2}{2}, \frac{100+W'}{2}\right), \tag{29}$$

$$m \,|\, (\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{D}_{-m}) \sim \mathrm{N}(RT, T), \tag{30}$$

where, IG denotes the inverse–gamma distribution and

$$W' = \sum_{i=0}^{2} \sum_{j=1}^{\bar{k}_i} \frac{(\tilde{M}^*_{j,i} - m)^2}{\tilde{V}^*_{j,i}},$$

$$T = \left[\frac{1}{2} + \frac{1}{\tau}\left(\sum_{i=1}^{\bar{k}_1} \frac{1}{\tilde{V}^*_{i,1}} + \sum_{j=1}^{\bar{k}_2} \frac{1}{\tilde{V}^*_{j,2}} + \sum_{r=1}^{\bar{k}_0} \frac{1}{\tilde{V}^*_{r,0}}\right)\right]^{-1},$$

$$R = \left[\frac{\bar{D}}{2} + \frac{1}{\tau}\left(\sum_{i=1}^{\bar{k}_1} \frac{\tilde{M}^*_{i,1}}{\tilde{V}^*_{i,1}} + \sum_{j=1}^{\bar{k}_2} \frac{\tilde{M}^*_{j,2}}{\tilde{V}^*_{j,2}} + \sum_{r=1}^{\bar{k}_0} \frac{\tilde{M}^*_{r,0}}{\tilde{V}^*_{r,0}}\right)\right]^{-1}.$$

Acceleration step

In order to speed up the mixing of the chain, at the end of every iteration, we resample the distinct values $\tilde{\theta}^*_{j,i}$, for $i = 0, 1, 2$ and $j = 1, \ldots, \bar{k}_i$, from their conditional distribution. This distribution depends on the choice of $\tilde{p}_1$ and $\tilde{p}_2$ only through their base measure P_0 and therefore it is the same for GM–$\mathscr{D}$ and GM–st models. For every $j = 1, \ldots, \tilde{k}_1$, the conditional distribution of $\tilde{\theta}^*_{j,1} = (\tilde{M}^*_{j,1}, \tilde{V}^*_{j,1})$ is normal/inverse–gamma. More specifically, we can sample from

$$\tilde{V}^*_{j,1} \sim \mathrm{IG}\left(1 + \frac{\tilde{n}_{j,1}}{2}, 1 + \frac{W''}{2}\right),$$

$$\tilde{M}^*_{j,1} \,|\, \tilde{V}^*_{j,1} \sim \mathrm{N}\left(\frac{m + \tau \sum_{(*)} x_{i,1}}{1 + \tau \tilde{n}_{j,1}}, \tilde{V}^*_{j,1} \frac{\tau}{1 + \tau \tilde{n}_{j,1}}\right),$$

where

$$W'' = \sum_{(*)} x^2_{i,1} + \frac{m^2 \tilde{n}_{j,1} - \sum_{(*)} x_{i,1} \left(2m + \tau \sum_{(*)} x_{i,1}\right)}{(1 + \tau \tilde{n}_{j,1})}$$

and $\sum_{(*)}$ denotes the sum over all the indexes corresponding to observations whose latent variable coincides with $\tilde{\theta}^*_{j,1}$. Analogous expressions, with obvious modifications, hold true for $\tilde{\theta}^*_{j,2}$ and $\tilde{\theta}^*_{j,0}$.